DOCUMENT RESUME

ED 458 254 TM 033 416

AUTHOR	Rudner, Lawrence M., Ed.; Schafer, William D., Ed.
TITLE	Practical Assessment, Research & Evaluation, 2000-2001.
INSTITUTION	ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.; Maryland Univ., College Park. Dept. of Measurement, Statistics & Evaluation.
ISSN	ISSN-1531-7714
PUB DATE	2001-00-00
NOTE	203p.; "Practical Assessment, Research & Evaluation" is an electronic-only journal covered on an article-by-article basis in "Current Index to Journals in Education" (CIJE). For the first 10 articles in Volume 7, see TM 523 914-933 (in CIJE). For articles 11-22 in Volume 7, see TM 033 417 and TM 523 934-945 (in CIJE).
AVAILABLE FROM	For full text: http://ericae.net/pare.
PUB TYPE	Collected Works - Serials (022) -- ERIC Publications (071)
JOURNAL CIT	Practical Assessment, Research & Evaluation; v7 2000-2001
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	*Educational Assessment; Educational Research; Elementary Secondary Education; *Evaluation Methods; Higher Education; Limited English Speaking; Portfolio Assessment; Program Evaluation; Regression (Statistics); Reliability; *Research Methodology; *Scoring Rubrics; Standards; *Student Evaluation; Validity
IDENTIFIERS	*Hierarchical Linear Modeling

ABSTRACT

This document consists of papers published in the electronic journal "Practical Assessment, Research & Evaluation" during 2000-2001: (1) "Advantages of Hierarchical Linear Modeling" (Jason W. Osborne); (2) "Prediction in Multiple Regression" (Jason W. Osborne); (3) Scoring Rubrics: What, When, and How?" (Barbara M. Moskal); (4) "Organizational Issues Related to Portfolio Assessment Implementation in the Classroom" (Renee Forgette-Giroux and Marielle Simon); (5) "Summarizing Change in Test Scores: Shortcomings of Three Common Methods" (Michael Russell); (6) "Using Expected Growth Size To Summarize Test Score Changes" (Michael Russell); (7) "Using Critical Thinking To Conduct Effective Searches of Online Resources" (Sarah K. Brem and Andrea J. Boyes); (8) "Fundamental Assessment Principles for Teachers and School Administrators" (James H. McMillan); (9) "A Framework for Developing an Effective Instructional Program for Limited English Proficient Students with Limited Formal Schooling" (Angelo Alcala); (10) "Scoring Rubric Development: Validity and Reliability" (Barbara M. Moskal and Jon A. Leyden); (11) "Assessments and Accountability" (Condensed Verstion) (Robert L. Linn); (12) "Self-ReportedGPA and SAT: A Methodological Note (Jerrell C. Cassady); (13) "Using State Standards and Tests To Improve Instruction" (Christopher Tienken and Michael Wilson); (14) "Computing the Expected Proportions of Misclassified Examinees" (Lawrence M. Rudner); (15) "Replication in Field Research" (Willoiam D. Schafer); (16) "Profile Analysis: Multidimensional Scaling Approach" (Cody S. Ding); (17) "An Overview of Content Analysis" (Steve Stemler); (18) "A Rubic for Scoring Postsecondary Academic Skills" (Marielle Simon and Renee Forgette-Giroux); (19) "Conducting Web-based Surveys" (David J. Solomon); (20) "The Stability of Undergraduate Students' Cognitive Test Anxiety Levels" (Jerrell C. Cassady); (21) "Alignment of

Reproductions supplied by EDRS are the best that can be made
from the original document.

Standards and Assessments as an Accountability Criterion" (Paul M. La Marca); and (22) "A Confirmatory Analysis of the Wechsler Adult Intelligence Scale-Third Edition: Is the Verbal/Performance Discrepancy Justified?" (Gordon E. Tabu). (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 458 254

Practical Assessment, Research & Evaluation, 2000-2001

Volume 7

Lawrence M. Rudner and William D. Schafer, Editors

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

● Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 033 416

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Editorial Board

Lawrence M. Rudner,
Univ of Maryland, co-Editor

William D. Schafer,
Univ of Maryland, co-Editor

Board of Editors

Kathryn Alvestad, Calvert County (MD) Public Schools

Filip Dochy,
Univ of Maastricht

Kurt F. Geisinger,
St. Thomas University

Gene V Glass,
Arizona State Univ

Arlen R. Gullickson
Western Michigan University

Robin K. Henson,
University of North Texas

Robert Marzano, Mid-continent Research for Education and Learning

M. Kevin Matter
Cherry Creek (C) Schools

Donna Mertens
Gallaudet University

Denise McKeon,
National Education Assoc

Joe O'Reilly,

Practical Assessment, Research and Evaluation (PARE) is an on-line journal published by the ERIC Clearinghouse on Assessment and Evaluation (ERIC/AE) and the Department of Measurement, Statistics, and Evaluation at the University of Maryland, College Park. Its purpose is to provide education professionals access to refereed articles that can have a positive impact on assessment, research, evaluation, and teaching practice, especially at the local education agency (LEA) level.

Manuscripts published in *Practical Assessment, Research and Evaluation* are scholarly syntheses of research and ideas about issues and practices in education. They are designed to help members of the community keep up-to-date with effective methods, trends and research developments. While they are most often prepared for practitioners, such as teachers, administrators, and assessment personnel who work in schools and school systems, *PARE* articles can target other audiences, including researchers, policy makers, parents, and students.

Manuscripts to be considered for *Practical Assessment, Research and Evaluation*

Mesa (AZ) Public Schools

Albert Oosterhof
Florida State University

Jason W. Osborne,
North Carolina State Univ

Michael Russell,
Boston College

Bruce Thompson,
Texas A&M

Marie Miller-Whitehead
Consultant

Lynn Winters, Long Beach
Unified Public Schools

Steven L. Wise,
James Madison University

should be short, 1500-2000 words or about four pages in length, exclusive of tables and references. They should conform to the stylistic conventions of the American Psychological Association (APA). See the Policies section of this web site for technical specifications and a list of suggested topics. Manuscripts should be submitted electronically to pare2@ericae.net. Articles appearing in *Practical Assessment, Research and Evaluation* also become available in the ERIC database through the ERIC Digest Series. Many articles published in *PARE* were previously published as part of the *ERIC/AE Digest Series.*

Permission is granted to distribute any article in this journal for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used.

Practical Assessment, Research and Evaluation is listed amoung the ejournals in education at the website for the AERA SIG "Communications Among Researchers".

The Educational Resources Information Center (ERIC) Clearinghouse on Assessment and Evaluation is a project of the National Library of Education, US Department of Education directed by the Department of Measurement, Statistics, and Evaluation at the University of Maryland in College Park.

Full-text Library | Search ERIC | Test Locator | ERIC System | Assessment Resources Calls for papers |
About us | Site map | Search | Help | Home

©1999 ERIC Clearinghouse on Assessment and Evaluation. All rights reserved.

Practical Assessmen Research & Evaluation

Volume 7, 2000 **Articles 1-10**

Article Pages

1 *Jason W. Osborne*: Advantages of Hierarchical Linear Modeling 7

2 *Jason W. Osborne:* Prediction in Multiple Regression 9

3 *Barbara M. Moskal*: Scoring Rubrics: What when and How? 9

4 *Renée Forgette-Giroux & Marielle Simon*: Organizational Issues Related to Portfolio Assessment Implementation in the Classroom 7

5 *Michael Russell*: Summarizing Change in Test Scores: Shortcomings of Three Common Methods 6

6 *Michael Russell*: Using Expected Growth Size to Summarize Test Score Changes 7

7 *Sarah K. Brem & Andrea J. Boyes*: Using Critical Thinking to Conduct Effective Searches of Online Resources 7

8 *James H. McMillan*: Fundamental Assessment Principles for Teachers and School Administrators 8

9 *Angelo Alcala*: A Framework for Developing an Effective Instructional Program for Limited English Proficient Students with Limited Formal Schooling 6

10 *Barbara M. Moskal & Jon A. Leydens*: Scoring Rubric Development: Validity and Reliability 11

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Osborne, Jason W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, 7(1). Available online: http://ericae.net/pare/getvn.asp?v=7&n=1. This paper has been viewed 5141 times since 1/10/00.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

Find articles in ERIC written by
Osborne, Jason W.

Advantages of Hierarchical Linear Modeling

Jason W. Osborne
University of Oklahoma

Hierarchical, or nested, data structures are common throughout many areas of research. However, until recently there has not been any appropriate technique for analyzing these types of data. Now, with several user-friendly software programs available, and some more readable texts and treatments on the topic, researchers need to be aware of the issue, and how it should be dealt with. The goal of this paper is to introduce the problem, how it is dealt with appropriately, and to provide examples of the pitfalls of not doing appropriate analyses.

What is a Hierarchical Data Structure?

People (and other living creatures, for that matter) tend to exist within organizational structures, such as families, schools, business organizations, churches, towns, states, and countries. In education, students exist within a hierarchical social structure that can include family, peer group, classroom, grade level, school, school district, state, and country. Workers exist within production or skill units, businesses, and sectors of the economy, as well as geographic regions. Health care workers and patients exist within households and families, medical practices and facilities (a doctor's practice, or hospital, e.g.), counties, states, and countries. Many other communities exhibit hierarchical data structures as well.

Bryk and Raudenbush (1992) also discuss two other types of data hierarchies that are less obvious: repeated-measures data and meta-analytic data. Once one begins looking for hierarchies in data, it becomes obvious that data repeatedly gathered on an individual is hierarchical, as all the observations are nested within individuals. While there are other adequate procedures for dealing with this sort of data, the assumptions relating to them are rigorous, whereas procedures relating to hierarchical modeling require fewer assumptions. Also, when researchers are engaged in the task of meta-analysis, or analysis of a large number of existing studies, it should become clear that subjects, results, procedures, and experimenters are nested within experiment. While this paper will not delve into these issues further, readers are encouraged to refer to Bryk and Raudenbush (1992) for further discussion of the advantages of hierarchical analysis for these types of data.

Why is a Hierarchical Data Structure an Issue?

Hierarchical, or nested, data present several problems for analysis. First, people or creatures that exist within hierarchies tend to be more similar to each other than people randomly sampled from the entire population. For example, students in a particular third-grade classroom are more similar to each other than to students randomly sampled from the school district as a whole, or from the national population of third-graders. This is because students are not randomly assigned to classrooms from the population, but rather are assigned to schools based on geographic factors. Thus, students within a particular classroom tend to come from a community or community segment that is more homogeneous in terms of morals and values, family background, socio-economic status, race or ethnicity, religion, and even educational preparation than the population as a whole. Further, students within a particular classroom share the experience of being in the same environment-- the same teacher, physical environment, and similar experiences, which may lead to increased homogeneity over time..

The problem of independence of observations. This discussion could be applied to any level of nesting, such as the family, the school district, county, state, or even country. Based on this discussion, we can assert that individuals who are drawn from an institution, such as a classroom, school, business, or health care unit, will be more homogeneous than if individuals were randomly sampled from a larger population. Herein lies the first issue for analysis of this sort of data. Because these individuals tend to share certain characteristics (environmental, background, experiential, demographic, or otherwise), observations based on these individuals are not fully independent. However, most analytic techniques require independence of observations as a primary assumption for the analysis. Because this assumption is violated in the presence of hierarchical data, ordinary least squares regression produces standard errors that are too small (unless these so-called design effects are incorporated into the analysis). In turn, this leads to a higher probability of rejection of a null hypothesis than if: (a) an appropriate statistical analysis were performed, or (b) the data included truly independent observations.

The problem of how to deal with cross-level data. Going back to the example of our third-grade classroom, it is often the case that a researcher is interested in understanding how environmental variables (e.g., teaching style, teacher behaviors, class size, class composition, district policies or funding, or even state or national variables, etc.) affect individual outcomes (e.g., achievement, attitudes, retention, etc.). But given that outcomes are gathered at the individual level, and other variables at classroom, school, district, state, or nation level, the question arises as to what the unit of analysis should be, and how to deal with the cross-level nature of the data.

One strategy would be to assign classroom or teacher (or school, district, or other) characteristics to all students (i.e., to bring the higher-level variables down to the student level). The problem with this approach, again, is non-independence of observations, as all students within a particular classroom assume identical scores on a variable.

Another way to deal with this issue would be to aggregate up to the level of the classroom, school, district, etc. Thus, we could talk about the effect of teacher or classroom characteristics on average classroom achievement. However, there are several issues with this approach, including: (a) that much (up to 80-90%) of the individual variability on the outcome variable is lost, which can lead to dramatic under- or over-estimation of observed relationships between variables (Bryk & Raudenbush, 1992), and (b) the outcome variable changes significantly and substantively from individual achievement to average classroom achievement.

Aside from these problems, both these strategies prevent the researcher from disentangling individual and group effects on the outcome of interest. As neither one of these approaches is satisfactory, the third approach, that of hierarchical modeling, becomes necessary.

How do Hierarchical Models Work? A Brief Primer

The goal of this paper is to introduce the concept of hierarchical modeling, and explicate the need for the procedure. It cannot fully communicate the nuances and procedures needed to actually perform a hierarchical analysis. The reader is encouraged to refer to Bryk and Raudenbush (1992) and the other suggested readings for a full explanation of the conceptual and methodological details of hierarchical modeling.

The basic concept behind hierarchical modeling is similar to that of OLS regression. On the base level (usually the individual level, referred to here as level 1), the analysis is similar to that of OLS regression: an outcome variable is predicted as a function of a linear combination of one or more level 1 variables, plus an intercept, as so:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_1 + \ldots + \beta_{kj}X_k + r_{ij}$$

where β_{0j} represents the intercept of group j, β_{1j} represents the slope of variable X_1 of group j, and r_{ij} represents the residual for individual i within group j. On subsequent levels, the level 1 slope(s) and intercept become dependent variables being predicted from level 2 variables:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_1 + \ldots + \gamma_{0k}W_k + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_1 + \ldots + \gamma_{1k}W_k + u_{1j}$$

and so forth, where γ_{00} and γ_{10} are intercepts, and γ_{01} and γ_{11} represent slopes predicting β_{0j} and β_{1j} respectively from variable W_1. Through this process, we accurately model the effects of level 1 variables on the outcome, and the effects of level 2 variables on the outcome. In addition, as we are predicting slopes as well as intercepts (means), we can model cross-level interactions, whereby we can attempt to understand what explains differences in the relationship between level 1 variables and the outcome. This will be discussed a bit more below.

An Empirical Comparison of the Three Approaches to Analyzing Hierarchical Data

To illustrate the outcomes achieved by each of the three possible analytic strategies for dealing with hierarchical data, disaggregation (bringing level 2 data down to level 1), aggregation, and multilevel modeling, data were drawn from the National Education Longitudinal Survey of 1988. This data set contains data on a representative sample of approximately 28,000 eighth graders in the United States at a variety of levels, including individual, family, teacher, and school. The analysis we performed predicted composite achievement test scores (math, reading combined) from student socioeconomic status (family SES), student locus of control (LOCUS), the percent of students in the school who are members of racial or ethnic minority groups (%MINORITY), and the percent of students in a school who receive free lunch (%LUNCH). Achievement is our outcome, SES and LOCUS are level 1 predictors, and %MINORITY and %LUNCH are level 2 indicators of school environment. In general, SES and LOCUS are expected to be positively related to achievement, and %MINORITY and %LUNCH are expected to be negatively related to achievement. In these analyses, 995 of a possible 1004 schools were represented (the remaining nine were removed due to insufficient data).

Disaggregated analysis. In order to perform the disaggregated analysis, the level 2 values were assigned to all individual students within a particular school (which is how the NELS data set comes). A standard multiple regression was performed via SPSS entering all predictor variables simultaneously. The resulting model was significant, with R=.56, R-square=.32, $\underline{F}$ (4,22899)=2648.54, $\underline{p} < .0001$. The individual regression weights and significance tests are presented in Table 1.

Table 1
Comparison of three analytic strategies.

	Disaggregated			Aggregated			Hierarchical		
Variable:	**B**	**SE**	**t**	**B**	**SE**	**t**	**B**	**SE**	**t**
SES	4.97_a	.08	62.11**	7.28_b	.26	27.91**	4.07_c	.10	41.29**
LOCUS	2.96_a	.08	37.71**	4.97_b	.49	10.22**	2.82_a	.08	35.74**
%MINORITY	-0.45_a	.03	-15.53**	-0.40_a	.06	-8.76**	-0.59_b	.07	-8.73**
%LUNCH	-0.43_a	.03	-13.50**	0.03_b	.05	0.59	-1.32_c	.07	-19.17**

Note: B refers to an unstandardized regression coefficient, and is used for the HLM analysis to represent the unstandardized regression coefficients produced therein, even though these are commonly labeled as betas and gammas. SE refers to standard error. Bs with different subscripts were found to be significantly different from other Bs within the row at $p < .05$. ** $p < .0001$.

All four variables were significant predictors of student achievement. As expected, SES and LOCUS were positively related to achievement, while %MINORITY and %LUNCH were negatively related.

Aggregated analysis. In order to perform the aggregated analysis, all level 1 variables (achievement, LOCUS, SES) were aggregated up to the school level (level 2) by averaging. A standard multiple regression was performed via SPSS entering all predictor variables simultaneously. The resulting model was significant, with R=.87, R-square=.75, $F(4,999)=746.41$, $p < .0001$. As seen in Table 1, both average SES and average LOCUS were significantly positively related to achievement, and %MINORITY was negatively related. In this analysis, %LUNCH was not a significant predictor of average achievement.

Multilevel analysis. In order to perform the multilevel analysis, a true multilevel analysis was performed via HLM, in which the respective level 1 and level 2 variables were specified appropriately. Note also that all level 1 predictors were centered at the group mean, and all level 2 predictors were centered at the grand mean. The resulting model demonstrated goodness of fit (Chi-square for change in model fit =4231.39, 5 df, $p < .0001$). this analysis reveals significant positive relationships between achievement and the level 1 predictors (SES and LOCUS), and strong negative relationships between achievement and the level 2 predictors (%MINORITY and %LUNCH). Further, the analysis revealed significant interactions between SES and both level 2 predictors, indicating that the slope for SES gets weaker as %LUNCH and as %MINORITY increases. Also, there was an interaction between LOCUS and %MINORITY, indicating that as %MINORITY increases, the slope for LOCUS weakens. There is no clearly equivalent analogue to R and R-square available in HLM.

Comparison of the Three Analytic Strategies and Conclusions

For the purposes of this discussion, we will assume that the third analysis represents the best estimate of what the "true" relationships are between the predictors and the outcome. Unstandardized regression coefficients (Bs in OLS, betas and gammas in HLM) were compared statistically via procedures outlined in Cohen and Cohen (1983).

In examining what is probably the most common analytic strategy for dealing with data such as these, the disaggregated analysis provided the best estimates of the level 1 effects in an OLS analysis. However, it significantly overestimated the effect of SES, and significantly and substantially underestimated the effects of the level 2 effects. The standard errors in this analysis are generally lower than they should be, particularly for the level 2 variables.

In comparison, the aggregated analysis overestimated the multiple correlation by more than 100%, overestimated the the regression slope for SES by 79% and for LOCUS by 76%, and underestimated the slopes for %MINORITY by 32% and for %LUNCH by 98%.

These analyses reveal the need for multilevel analysis of multilevel data. Neither OLS analysis accurately modeled the true relationships between the outcome and the predictors. Additionally, HLM analyses provide other benefits, such as easy modeling of cross-level interactions, which allows for more interesting questions to be asked of the data. With nested and hierarchical data common in the social and other sciences, and with recent developments making HLM software packages more user-friendly and accessible, it is important for researchers in all fields to become acquainted with these procedures.

ADDITIONAL READINGS

Bryk, A.S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20 (2), 115-147.

Hoffman, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, 24(5), 623-641.

Nezlek, J. B., & Zyzniewski, L. E. (1998). Using hierarchical linear modeling to analyze grouped data. *Group Dynamics*, 2, 313-320.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* (pp. 675-711). Harcourt Brace: Orlando, FL.

Raudenbush, S. W. (1995). Reexamining, reaffirming, and improving application of hierarchical models. *Journal of Educational and Behavioral Statistics*, 20 (2), 210-220.

AUTHOR NOTES

Correspondence relating to this article can be addressed to Jason W. Osborne, Department of Educational Psychology, University of Oklahoma, 820 Van Vleet Oval, Norman, OK, 73019, or via email at josborne@ou.edu. The author would like to express his appreciation to all the faculty at the University of Oklahoma, and authors around the world, that have inspired this article in one way or another. Additional thanks go to my wonderful students who inspired me to write this article in the first place. Finally, acknowledgements are also due to Anthony Bryk and Stephen Raudenbush, who provided a book from which I have drawn many of my ideas and arguments relating to this paper.

Descriptors: *Hierarchical Linear Modeling; Estimation; Research Design; Research Methodology

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Osborne, Jason W. (2000). Prediction in multiple regression. *Practical Assessment, Research & Evaluation*, 7(2). Available online: http://ericae.net/pare/getvn.asp?v=7&n=2. This paper has been viewed 3501 times since 3/10/00.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

Find articles in ERIC written by
Osborne, Jason W.

Prediction in Multiple Regression

Jason W. Osborne
University of Oklahoma

There are two general applications for multiple regression (MR): prediction and explanation[1]. These roughly correspond to two differing goals in research: being able to make valid projections concerning an outcome for a particular individual (prediction), or attempting to understand a phenomenon by examining a variable's correlates on a group level (explanation). There has been debate as to whether these two applications of MR are grossly different, as Scriven (1959) and Anderson and Shanteau, (1977) asserts, or necessarily part and parcel of the same process (e.g., DeGroot, 1969, Kaplan, 1964; for an overview of this discussion see Pedhazur, 1997, pp. 195-198). Regardless of the philosophical issues, there are different analytic procedures involved with the two types of analyses. The goal of this paper is to present: (a) the concept of prediction via MR, (b) the assumptions underlying multiple regression analysis, (c) shrinkage, cross-validation, and double cross-validation of prediction equations, and (d) how to calculate confidence intervals around individual predictions.

What is the Difference between Using MR for Prediction versus Using MR for Explanation?

When one uses MR for explanatory purposes, that person is exploring relationships between multiple variables in a sample to shed light on a phenomenon, with a goal of generalizing this new understanding to a population. When one uses MR for prediction, one is using a sample to create a regression equation that would optimally predict a particular phenomenon within a particular population. Here the goal is to use the equation to predict outcomes for individuals *not in the sample used in the analysis*. Hypothetically, researchers might create a regression equation to predict twelfth-grade achievement test scores from eighth-grade variables, such as family socioeconomic status, race, sex, educational plans, parental education, GPA, and participation in school-based extracurricular activities. The goal is not to understand why students achieve at a certain level, but to create the best equation so that, for example, guidance counselors could predict future achievement scores for their students, and (hopefully) intervene with those students identified as at risk for poor performance, or to select students into programs based on their projected scores. And while theory is useful for identifying what variables should be in a prediction equation, the variables do not necessarily need to make conceptual sense. If the single greatest predictor of future achievement scores was the number of hamburgers a student eats, it should be in the prediction equation regardless of whether it makes sense (although this sort of finding might spur some explanatory research....)

How is a Prediction Equation Created?

The general process for creating a prediction equation involves gathering relevant data from a large, representative sample from the population. What constitutes "large" is open to debate, and while guidelines for general applications of regression are as small as 50 + 8*number of predictors (Tabachnick & Fidell, 1996), guidelines for prediction equations are more stringent due to the need to generalize beyond a given sample. While some authors have suggested that 15 subjects per predictor is sufficient (Park & Dudycha, 1974; Pedhazur, 1997), others have suggested minimum total sample (e.g., 400, see Pedhazur, 1997), others have suggested a minimum of 40 subjects per predictor (Cohen and Cohen, 1983; Tabachnick & Fidell, 1996). Of course, as the goal is a stable regression equation that is representative of the population regression equation, more is better. If one has good estimates of effect sizes, a power analysis might give a good estimate of the sample size. The effect of sample size on shrinkage and stability will be explored below.

Methods for entering variables into the equation. There are many ways to enter predictors into the regression equation. Several of these rely on the statistical properties of the variables to determine order of entry (e.g., forward selection, backward elimination, stepwise). Others rely on the experimenter to specify order of entry (hierarchical, blockwise), or have no order of entry (simultaneous). Current practice clearly favors analyst-controlled entry, and discourages entry based on the statistical properties of the variables as it is atheoretical. A thorough discussion of this issue is beyond the scope of this paper, so the reader is referred to Cohen and Cohen (1983) and Pedhazur (1997) for overviews of the various techniques, and to Thompson (1989) and Schafer (1991a, 1991b) for more detailed discussions of the issues.

Regardless of the method ultimately chosen by the researcher, it is critical that the researcher examine individual variables to ensure that only variables contributing significantly to the variance accounted for by the regression equation are included. Variables not accounting for significant portions of variance should be deleted from the equation, and the equation should be re-calculated. Further, researchers might want to examine excluded variables to see if their entry would significantly improve prediction (a significant increase in R-squared).

What Assumptions Must be Met When Doing a Regression Analysis?

It is absolutely critical that researchers assess whether their analyses meet the assumptions of multiple regression. These assumptions are explained in detail in places such as Pedhazur (1997) and Cohen and Cohen (1983), and as such will not be addressed further here. Failure to meet necessary assumptions can cause problems with prediction equations, often serving to either make them less generalizable than they otherwise would be, or causing underprediction (accounting for less variance than they should, such as in the case of curvilinearity or poor measurement).

How Are Prediction Equations Evaluated?

In a prediction analysis, the computer will produce a regression equation that is optimized for the sample. Because this process capitalizes on chance and error in the sample, the equation produced in one sample will not generally fare as well in another sample (i.e., R-squared in a subsequent sample using the same equation will not be as large as R-squared from original sample), a phenomenon called shrinkage. The most desirable outcome in this process is for minimal shrinkage, indicating that the prediction equation will generalize well to new samples or individuals from the population examined. While there are equations that can estimate shrinkage, the best way to estimate shrinkage, and test the prediction equation is through cross-validation or double-cross validation.

Cross-validation. To perform cross-validation, a researcher will either gather two large samples, or one very large sample which will be split into two samples via random selection procedures. The prediction equation is created in the first sample. That equation is then used to create predicted scores for the members of the second sample. The predicted scores are then correlated with the observed scores on the dependent variable ($\mathbf{r}_{yy'}$). This is called the *cross-validity coefficient*. The difference between the original R-squared and $\mathbf{r}_{yy'}^2$ is the shrinkage. The smaller the shrinkage, the more confidence we can have in the generalizability of the equation.

In our example of predicting twelfth-grade achievement test scores from eighth-grade variables a sample of 700 students (a subset of the larger National Education Longitudinal Survey of 1988) were randomly split into two groups. In the first group, analyses revealed that the following eighth-grade variables were significant predictors of twelfth-grade achievement: GPA, parent education level, race (white=0, nonwhite=1), and participation in school-based extracurricular activities (no=0, yes=1), producing the following equation:

$$Y' = -2.45 + 1.83(\text{GPA}) - 0.77(\text{Race}) + 1.03(\text{Participation}) + 0.38(\text{Parent Ed})$$

In the first group, this analyses produced an R-squared of .55. This equation was used in the second group to create predicted scores, and those predicted scores correlated $\mathbf{r}_{yy'} = .73$ with observed achievement scores. With a $\mathbf{r}_{yy'}^2$ of .53 (cross-validity coefficient), shrinkage was 2%, a good outcome.

Double cross-validation. In double cross-validation prediction equations are created in both samples, and then each is used to create predicted scores and cross-validity coefficients in the other sample. This procedure involves little work beyond cross-validation, and produces a more informative and rigorous test of the generalizability of the regression equation(s). Additionally, as two equations are produced, one can look at the stability of the actual regression line equations.

The following regression equation emerged from analyses of the second sample::

$$Y' = -4.03 + 2.16(\text{GPA}) - 1.90(\text{Race}) + 1.43(\text{Participation}) + 0.28(\text{Parent Ed})$$

This analysis produced an R-squared of .60. This equation was used in the first group to create predicted scores in the first group, which correlated .73 with observed scores, for a cross-validity coefficient of .53. Note that: (a) the second analysis revealed larger shrinkage than the first, (b) the two cross-validation coefficients were identical (.53), and (c) the two regression equations are markedly different, even though the samples had large subject to predictor ratios (over 80:1).

How much shrinkage is too much shrinkage? There are no clear guidelines concerning how to evaluate shrinkage, except the general agreement that less is always better. But is 3% acceptable? What about 5%? 10%? Or should it be a proportion of the original R-squared (so that 5% shrinkage on an R-squared of .50 would be fine, but 5% shrinkage on an R-squared of .30 would not be)? There are no guidelines in the literature. However, Pedhazur has suggested that one of the advantages of double cross-validation is that one can compare the two cross-validity coefficients, and if similar, one can be fairly confident in the generalizability of the equation.

The final step. If you are satisfied with your shrinkage statistics, the final step in this sort of analysis is to combine both samples (assuming shrinkage is minimal) and create a final prediction equation based on the larger sample. In our data set, the combined sample produced the following regression line equation:

$$Y' = -3.23 + 2.00(\text{GPA}) - 1.29(\text{Race}) + 1.24(\text{Participation}) + 0.32(\text{Parent Ed})$$

How does sample size affect the shrinkage and stability of a prediction equation?

As discussed above, there are many different opinions as to the minimum sample size one should use in prediction research. As an illustration of the effects of different subject to predictor ratios on shrinkage and stability of a regression equation, data from the National Education Longitudinal Survey of 1988 (NELS 88, from the National Center for Educational Statistics) were used to construct prediction equations identical to our running example. This data set contains data on 24,599 eighth grade students representing 1052 schools in the United States. Further, the data can be weighted to exactly represent the population, so an accurate population estimate can be obtained for comparison. Two samples, each representing ratios of 5, 15, 40, 100, and 400 subjects per predictor were randomly selected from this sample (randomly selecting from the full sample for each new pair of a different size). Following selection of the samples, prediction equations were calculated, and double cross-validation was performed. The results are presented in Table 1.

Table 1: Comparison of double cross validation results
with differing subject:predictor ratios

Sample Ratio (subjects: predictors)	**Obtained Prediction Equation**	R^2	$r_{yy'}^2$	**Shrink-age**
Population	Y'= -1.71+2.08(GPA) -0.73(race) -0.60(part) +0.32(pared)	.48		
5:1				
Sample 1	Y'= -8.47 +1.87(GPA) -0.32(race) +5.71(part) +0.28(pared)	.62	.53	.09
Sample 2	Y'= -6.92 +3.03(GPA) +0.34(race) +2.49 (part) -0.32(pared)	.81	.67	.14
15:1				
Sample 1	Y'= -4.46 +2.62(GPA) -0.31(race) +0.30(part) +0.32(pared)	.69	.24	.45
Sample 2	Y'= -1.99 +1.55(GPA) +0.34(race) +1.04 (part) -0.58(pared)	.53	.49	.04
40:1				
Sample 1	Y'= -0.49 +2.34(GPA) -0.79(race) -1.51(part) +0.08(pared)	.55	.50	.05
Sample 2	Y'= -2.05 +2.03(GPA) -0.61(race) -0.37(part) -+0.51(pared)	.58	.53	.05

100:1				
Sample 1	Y'= -1.89 +2.05(GPA) -0.52(race) - 0.17(part) +0.35(pared)	.46	.45	.01
Sample 2	Y'= -2.04 +1.92(GPA) -0.01(race) +0.32(part) +0.37(pared)	.46	.45	.01
400:1				
Sample 1	Y'= -1.26 +1.95(GPA) -0.70(race) - 0.41(part) +0.37(pared)	.47	.46	.01
Sample 2	Y'= -1.10 +1.94(GPA) -0.45(race) - 0.56(part) +0.35(pared)	.42	.41	.01

The first observation from the table is that, by comparing regression line equations, the very small samples have wildly fluctuating equations (both intercept and regression coefficients). Even the 40:1 ratio samples have impressive fluctuations in the actual equation. While the fluctuations in the 100:1 sample are fairly small in magnitude, some coefficients reverse direction, or are far off of the population regression line. As expected, it is only in the largest ratios presented, the 100:1 and 400:1 ratios, that the equations stabilize and remain close to the population equation.

Comparing variance accounted for, variance accounted for is overestimated in the equations with less than a 100:1 ratio. Cross-validity coefficients vary a great deal across samples until a 40:1 ratio is reached, where they appear to stabilize. Finally, it appears that shrinkage appears to minimize as a 40:1 ratio is reached. If one takes Pedhazur's suggestion to compare cross-validity coefficients to determine if your equation is stable, from these data one would need a 40:1 ratio or better before that criterion would be reached. If the goal is to get an accurate, stable estimate of the population regression equation (which it should be if that equation is going to be widely used outside the original sample), it appears desirable to have at least 100 subjects per predictor.

Calculating a Predicted Score, and Confidence Intervals Around That Score

There are two categories of predicted scores relevant here: scores predicted for the original sample, and scores that can be predicted for individuals outside the original sample. Individual predicted scores and confidence intervals for the original sample are available in the output available from most common statistical packages. Thus, the latter will be addressed here.

Once an analysis is completed and the final regression line equation is formed, it is possible to create predictions for individuals who were not part of the original sample that generated the regression line (one of the attractive features of regression). Calculating a new score based on an existing regression line is a simple matter of substitution and algebra. However, no such prediction should be presented without confidence intervals. The only practical way to do this is through the following formula:

$$Y' \pm t_{(\alpha/2,\, df)}\,(s_{y'})$$

where $s_{y'}$ is calculated as: $\sqrt{S_{\mu'}^{2} + MS_{residual}}$

where $S_{\mu'}^{2}$ is the squared standard error of mean predicted scores (standard error of the estimate, squared), and the mean square residual, both of which can be obtained from typical regression output.[2]

Summary and suggestions for further study

Multiple regression can be an effective tool for creating prediction equations providing adequate measurement, large enough samples, assumptions of MR are met, and care is taken to evaluate the regression equations for generalizability (shrinkage). Researchers interested in this topic might want to explore the following topics: (a) the use of logistic regression for predicting binomial or discrete outcomes, (b) the use of estimation procedures other than ordinary least squares regression that can produce better prediction (e.g., Bayesian estimation, see e.g. Bryk and Raudenbush, 1992), and (c) alternatives to MR when assumptions are not met, or when sample sizes are inadequate to produce stable estimates, such as ridge regression (for an introduction to these alternative procedures see e.g., Cohen & Cohen, 1983, pp.113-115). Finally, if researchers have nested or multilevel data, they should use multilevel modeling procedures (e.g., HLM, see Bryk & Raudenbush, 1992) to produce prediction equations.

SUGGESTED READING:

Anderson, N. H., & Shanteau, J. (1977). Weak inference with linear models. *Psychological Bulletin, 84,* 1155-1170.

Bryk, A.S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage Publications.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

DeGroot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences.* The Hague: Mouton.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* San Francisco: Chandler.

Park, C., & Dudycha, A. (1974). A cross-validation approach to sample size determination. *Journal of the American Statistical Association, 69,* 214-218.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research.* Harcourt Brace: Orlando, FL.

Scriven, M. (1959). Explanation and prediction in evolutionary theory. *Science, 130,* 477-482.

Thompson, B. (1989). Why won't stepwise methods die? *Measurement and Evaluation in Counseling and Development, 21,* 146-148.

Schafer, W. D. (1991a). Reporting hierarchical regression results. *Measurement and Evaluation in Counseling and Development, 24,* 98-100.

Schafer, W.D. (1991b). Reporting nonhierarchical regression results. *Measurement and Evaluation in Counseling and Development, 24,* 146-149.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using Multivariate Statistics.* New York: Harper Collins.

FOOTNOTES

1. Some readers may be uncomfortable with the term "explanation" when referring to multiple regression, as these data are often correlational in nature, while the term explanation often implies causal inference. However, explanation will be used in this article because: (a) it is the convention in the field, (b) here we are talking of regression with the *goal* of explanation, and (c) one can come to understanding of phenomena by understanding associations without positing or testing strict causal orderings..

2. It is often the case that one will want to use standard error of the predicted score when calculating an individual confidence interval. However, as that statistic is only available from statistical program output, and only for individuals in the original data set, it is of limited value for this discussion. Here we suggest using the standard error of the mean predicted scores, as it is the best estimate of the standard error of the predicted score, knowing it is not completely ideal, but lacking any other alternative.

AUTHOR NOTES

Correspondence relating to this article can be addressed to Jason W. Osborne, Department of Educational Psychology, University of Oklahoma, 820 Van Vleet Oval, Norman, OK, 73019, or via email at josborne@ou.edu. Special thanks go to William Schafer, whose many good suggestions and critical eye helped to substantially shape this paper.

Descriptors: Multiple Regression; Modeling; Prediction; Research Design; Research Methods

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Moskal, Barbara M. (2000). Scoring rubrics: what, when and how?. *Practical Assessment, Research & Evaluation*, 7(3). Available online: http://ericae.net/pare/getvn.asp?v=7&n=3. This paper has been viewed 21371 times since 3/29/00.

- Find similar papers in
  ERICAE Full Text Lit
  Pract Assess, Res &
  ERIC RIE & CIJE 19
  ERIC On-Demand D
  ERIC/AE Abstracts I
- Find articles in ERIC w
  Moskal, Barbara M.

Scoring Rubrics: What, When and How?

Barbara M. Moskal
Associate Director of the Center for Engineering Education
Assistant Professor of Mathematical and Computer Sciences
Colorado School of Mines

Scoring rubrics have become a common method for evaluating student work in both the K
the college classrooms. The purpose of this paper is to describe the different types of scori
rubrics, explain why scoring rubrics are useful and provide a process for developing scorir
rubrics. This paper concludes with a description of resources that contain examples of the
types of scoring rubrics and further guidance in the development process.

What is a scoring rubric?

Scoring rubrics are descriptive scoring schemes that are developed by teachers or other ev
to guide the analysis of the products or processes of students' efforts (Brookhart, 1999). Sc
rubrics are typically employed when a judgement of quality is required and may be used t
evaluate a broad range of subjects and activities. One common use of scoring rubrics is to
the evaluation of writing samples. Judgements concerning the quality of a given writing s
may vary depending upon the criteria established by the individual evaluator. One evalu
heavily weigh the evaluation process upon the linguistic structure, while another evaluat
be more interested in the persuasiveness of the argument. A high quality essay is likely t
combination of these and other factors. By developing a pre-defined scheme for the evalua
process, the subjectivity involved in evaluating an essay becomes more objective.

Figure 1 displays a scoring rubric that was developed to guide the evaluation of student w

samples in a college classroom (based loosely on Leydens & Thompson, 1997). This is an e
of a holistic scoring rubric with four score levels. Holistic rubrics will be discussed in deta
this document. As the example illustrates, each score category describes the characteristi
response that would receive the respective score. By having a description of the character
responses within each score category, the likelihood that two independent evaluators wou
the same score to a given response is increased. This concept of examining the extent to w
independent evaluators assign the same score to a given response is referred to as "rater
reliability."

Figure 1.

Example of a scoring rubric designed to evaluate college writing samples.

-3-

Meets Expectations for a first Draft of a Professional Report

- The document can be easily followed. A combination of the following are apparer the document:

1. Effective transitions are used throughout,
2. A professional format is used,
3. The graphics are descriptive and clearly support the document's purpose.

- The document is clear and concise and appropriate grammar is used throughout

-2-

Adequate

- The document can be easily followed. A combination of the following are apparer the document:

1. Basic transitions are used,
2. A structured format is used,
3. Some supporting graphics are provided, but are not clearly explained.

- The document contains minimal distractions that appear in a combination of thc following forms:

1. Flow in thought
2. Graphical presentations
3. Grammar/mechanics

-1-

Needs Improvement

- Organization of document is difficult to follow due to a combination of following:

1. Inadequate transitions
2. Rambling format
3. Insufficient or irrelevant information
4. Ambiguous graphics

- The document contains numerous distractions that appear in the a combination the following forms:

1. Flow in thought
2. Graphical presentations
3. Grammar/mechanics

-0-

Inadequate

- There appears to be no organization of the document's contents.
- Sentences are difficult to read and understand.

When are scoring rubrics an appropriate evaluation technique?

Writing samples are just one example of performances that may be evaluated using scori rubrics. Scoring rubrics have also been used to evaluate group activities, extended project oral presentations (e.g., Chicago Public Schools, 1999; Danielson, 1997a; 1997b; Schrock, Moskal, 2000). They are equally appropriate to the English, Mathematics and Science cla (e.g., Chicago Public Schools, 1999; State of Colorado, 1999; Danielson, 1997a; 1997b; Da Marquez, 1998; Schrock, 2000). Both pre-college and college instructors use scoring rubri classroom evaluation purposes (e.g., State of Colorado, 1999; Schrock, 2000; Moskal, 2000 Moskal & Pavelich, 2000). Where and when a scoring rubric is used does not depend on t level or subject, but rather on the purpose of the assessment.

Scoring rubrics are one of many alternatives available for evaluating student work. For e checklists may be used rather then scoring rubrics in the evaluation of writing samples. C are an appropriate choice for evaluation when the information that is sought is limited to determination of whether specific criteria have been met. Scoring rubrics are based on de scales and support the evaluation of the extent to which criteria has been met.

The assignment of numerical weights to sub-skills within a process is another evaluation technique that may be used to determine the extent to which given criteria has been met. Numerical values, however, do not provide students with an indication as to how to impr performance. A student who receives a "70" out of "100", may not know how to improve hi performance on the next assignment. Scoring rubrics respond to this concern by providin descriptions at each level as to what is expected. These descriptions assist the students in understanding why they received the score that they did and what they need to do to imp their future performances.

Whether a scoring rubric is an appropriate evaluation technique is dependent upon the p the assessment. Scoring rubrics provide at least two benefits in the evaluation process. Fi support the examination of the extent to which the specified criteria has been reached. Se they provide feedback to students concerning how to improve their performances. If these are consistent with the purpose of the assessment, than a scoring rubric is likely to be an appropriate evaluation technique.

What are the different types of scoring rubrics?

Several different types of scoring rubrics are available. Which variation of the scoring rub should be used in a given evaluation is also dependent upon the purpose of the evaluation section describes the differences between analytic and holistic scoring rubrics and betwee specific and general scoring rubrics.

Analytic verses Holistic

In the initial phases of developing a scoring rubric, the evaluator needs to determine wha the evaluation criteria. For example, two factors that may be considered in the evaluatio writing sample are whether appropriate grammar is used and the extent to which the giv argument is persuasive. An analytic scoring rubric, much like the checklist, allows for th separate evaluation of each of these factors. Each criterion is scored on a different descrip (Brookhart, 1999).

The rubric that is displayed in Figure 1 could be extended to include a separate set of crit the evaluation of the persuasiveness of the argument. This extension would result in an a scoring rubric with two factors, quality of written expression and persuasiveness of the a Each factor would receive a separate score. Occasionally, numerical weights are assigned evaluation of each criterion. As discussed earlier, the benefit of using a scoring rubric rat weighted scores is that scoring rubrics provide a description of what is expected at each s Students may use this information to improve their future performance.

Occasionally, it is not possible to separate an evaluation into independent factors. When t an overlap between the criteria set for the evaluation of the different factors, a holistic sc rubric may be preferable to an analytic scoring rubric. In a holistic scoring rubric, the crit considered in combination on a single descriptive scale (Brookhart, 1999). Holistic scorin support broader judgements concerning the quality of the process or the product.

Selecting to use an analytic scoring rubric does not eliminate the possibility of a holistic f holistic judgement may be built into an analytic scoring rubric as one of the score catego difficulty with this approach is that overlap between the criteria that is set for the holisti judgement and the other evaluated factors cannot be avoided. When one of the purposes o evaluation is to assign a grade, this overlap should be carefully considered and controlled evaluator should determine whether the overlap is resulting in certain criteria are being more than was originally intended. In other words, the evaluator needs to be careful that student is not unintentionally severely penalized for a given mistake.

General verses Task Specific

Scoring rubrics may be designed for the evaluation of a specific task or the evaluation of a category of tasks. If the purpose of a given course is to develop a student's oral communic skills, a general scoring rubric may be developed and used to evaluate each of the oral presentations given by that student. This approach would allow the students to use the fe that they acquired from the last presentation to improve their performance on the next presentation.

If each oral presentation focuses upon a different historical event and the purpose of the assessment is to evaluate the students' knowledge of the given event, a general scoring ru evaluating a sequence of presentations may not be adequate. Historical events differ in b influencing factors and outcomes. In order to evaluate the students' factual and conceptu knowledge of these events, it may be necessary to develop separate scoring rubrics for eac presentation. A "Task Specific" scoring rubric is designed to evaluate student performanc single assessment event.

Scoring rubrics may be designed to contain both general and task specific components. If purpose of a presentation is to evaluate students' oral presentation skills and their knowl the historical event that is being discussed, an analytic rubric could be used that contains general component and a task specific component. The oral component of the rubric may a general set of criteria developed for the evaluation of oral presentations; the task specifi component of the rubric may contain a set of criteria developed with the specific historica mind.

How are scoring rubrics developed?

The first step in developing a scoring rubric is to clearly identify the qualities that need t displayed in a student's work to demonstrate proficient performance (Brookhart, 1999). T identified qualities will form the top level or levels of scoring criteria for the scoring rubri decision can then be made as to whether the information that is desired from the evaluati best be acquired through the use of an analytic or holistic scoring rubric. If an analytic sc rubric is created, then each criterion is considered separately as the descriptions of the di score levels are developed. This process results in separate descriptive scoring schemes fo evaluation factor. For holistic scoring rubrics, the collection of criteria is considered throu the construction of each level of the scoring rubric and the result is a single descriptive sc scheme.

After defining the criteria for the top level of performance, the evaluator's attention may l to defining the criteria for lowest level of performance. What type of performance would s very limited understanding of the concepts that are being assessed? The contrast between criteria for top level performance and bottom level performance is likely to suggest appro criteria for middle level of performance. This approach would result in three score levels.

If greater distinctions are desired, then comparisons can be made between the criteria for existing score level. The contrast between levels is likely to suggest criteria that may be u create score levels that fall between the existing score levels. This comparison process car until the desired number of score levels is reached or until no further distinctions can be ı meaningful distinctions between the score categories cannot be made, then additional sco categories should not be created (Brookhart, 1999). It is better to have a few meaningful s categories then to have many score categories that are difficult or impossible to distinguis

Each score category should be defined using descriptions of the work rather then judgeme about the work (Brookhart, 1999). For example, "Student's mathematical calculations con errors," is preferable over, "Student's calculations are good." The phrase "are good" requir evaluator to make a judgement whereas the phrase "no errors" is quantifiable. In order tc determine whether a rubric provides adequate descriptions, another teacher may be aske the scoring rubric to evaluate a sub-set of student responses. Differences between the scoı assigned by the original rubric developer and the second scorer will suggest how the rubri further clarified.

Resources

Currently, there is a broad range of resources available to teachers who wish to use scorii in their classrooms. These resources differ both in the subject that they cover and the levı they are designed to assess. The examples provided below are only a small sample of the information that is available.

For K-12 teachers, the State of Colorado (1998) has developed an on-line set of general, h scoring rubrics that are designed for the evaluation of various writing assessments. The (Public Schools (1999) maintain an extensive electronic list of analytic and holistic scoring that span the broad array of subjects represented throughout K-12 education. For mather teachers, Danielson has developed a collection of reference books that contain scoring rub are appropriate to the elementary, middle school and high school mathematics classroom 1997b; Danielson & Marquez, 1998).

Resources are also available to assist college instructors who are interested in developing using scoring rubrics in their classrooms. *Kathy Schrock's Guide for Educators* (2000) con electronic materials for both the pre-college and the college classroom. In *The Art and Sci Classroom Assessment: The Missing Part of Pedagogy*, Brookhart (1999) provides a brief, l comprehensive review of the literature on assessment in the college classroom. This inclu description of scoring rubrics and why their use is increasing in the college classroom. Mo (1999) has developed a web site that contains links to a variety of college assessment reso including scoring rubrics.

The resources described above represent only a fraction of those that are available. The E Clearinghouse on Assessment and Evaluation [ERIC/AE] provides several additional usel sites. One of these, *Scoring Rubrics - Definitions & Constructions* (2000b), specifically add questions that are frequently asked with regard to scoring rubrics. This site also provides electronic links to web resources and bibliographic references to books and articles that di scoring rubrics. For more recent developments within assessment and evaluation, a searc completed on the abstracts of papers that will soon be available through ERIC/AE (2000a site also contains a direct link to ERIC/AE abstracts that are specific to scoring rubrics.

Search engines that are available on the web may be used to locate additional electronic When using this approach, the search criteria should be as specific as possible. Generic se that use the terms "rubrics" or "scoring rubrics" will yield a large volume of references. W seeking information on scoring rubrics from the web, it is advisable to use an advanced se specify the grade level, subject area and topic of interest. If more resources are desired th from this conservative approach, the search criteria can be expanded.

References

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: T George Washington University, Graduate School of Education and Human Developr

Chicago Public Schools (1999). *Rubric Bank*. [Available online at: http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Rubric_Bank/rubric_ba

Danielson, C. (1997a). *A Collection of Performance Tasks and Rubrics: Middle Schoo Mathematics*. Larchmont, NY: Eye on Education Inc.

Danielson, C. (1997b). *A Collection of Performance Tasks and Rubrics: Upper Eleme School Mathematics*. Larchmont, NY: Eye on Education Inc.

Danielson, C. & Marquez, E. (1998). *A Collection of Performance Tasks and Rubrics: School Mathematics*. Larchmont, NY: Eye on Education Inc.

ERIC/AE (2000a). *Search ERIC/AE draft abstracts*. [Available online at: http://ericae.net/sinprog.htm].

ERIC/AE (2000b). *Scoring Rubrics - Definitions & Construction* [Available online at: http://ericae.net/faqs/rubrics/scoring_rubrics.htm].

Knecht, R., Moskal, B. & Pavelich, M. (2000). *The Design Report Rubric: Measuring Tracking Growth through Success*, Paper to be presented at the annual meeting of tl American Society for Engineering Education.

Leydens, J. & Thompson, D. (August, 1997), *Writing Rubrics Design (EPICS) I*, Inte Communication, Design (EPICS) Program, Colorado School of Mines.

Moskal, B. (2000). *Assessment Resource Page*. [Available online at: http://www.mines.edu/Academic/assess/Resource.htm].

Schrock, K. (2000). *Kathy Schrock's Guide for Educators*. [Available online at: http://school.discovery.com/schrockguide/assess.html].

State of Colorado (1998). The Rubric. [Available online at: http://www.cde.state.co.us/cdedepcom/asrubric.htm#writing].

Descriptors: *Rubrics; Scoring; *Student Evaluation; *Test Construction; *Evaluation Methods; Grades; Grading; *Scoring

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Forgette-Giroux, Renée & Marielle Simon (2000). Organizational issues related to portfolio assessment implementation in the classroom. *Practical Assessment, Research & Evaluation*, 7(4). Available online: http://ericae.net/pare/getvn.asp?v=7&n=4. This paper has been viewed 12299 times since 4/19/00.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Forgette-Giroux, Renée
  Marielle Simon

Organizational Issues Related to Portfolio Assessment Implementation in the Classroom[1]

Renée Forgette-Giroux & Marielle Simon
Faculty of Education, University of Ottawa

This paper explores organizational issues that arose when implementing portfolio assessment in eleven classrooms during the field trial of a generic content selection framework[2]. Some researchers have already examined, to various degrees, the organizational process teachers go through when implementing portfolios within their classrooms to assess learning as opposed to showcasing or reporting achievement. Their results point to four categories of factors that seemed to affect portfolio implementation process:

1. the need for frameworks to guide entry selection and interpretation (Smit, Kolonosky, & Seltzer 1991),
2. teacher training or development (Smit, et. al., 1991),
3. time (Glazer, 1994; Lescher, 1998) and
4. teaching styles and values (Sawyer, 1994).

Other studies have suggested the existence of a possible portfolio assessment implementation process or continuum. Fingeret's (1993) study, for example, led to the identification of a four-stage process revolving around specific tasks such as examining fit within assessment practices to revision of actual portfolio use. Calkins (1992), on the other hand, ranked teachers on a five-point continuum based on their level of acceptance of the portfolio and its integration within their teaching styles and approaches. The purpose of this paper is to highlight possible relationships between input and process variables and resulting organizational issues surrounding portfolio assessment implementation in the classroom when a generic content selection framework is provided.

Content selection framework

Portfolio assessment is defined here as *a cumulative and ongoing collection of entries that are selected and commented on by the student, the teacher and/or peers, to assess the student's progress in the development of a competency* (Simon, & Forgette-Giroux, 2000). The generic portfolio assessment content selection framework recommends the collection of entries (items or contents) along five learning dimensions of a competency:

a. cognitive,
b. affective,
c. behavioural,
d. metacognitive, and
e. developmental.

The pieces of evidence are combined to provide an interrelated, complete, dynamic, and holistic picture of the students' development toward mastery of a complex skill such as problem solving or oral communication. Whereas the five categories are considered fixed within the framework, organizational decisions regarding storing, scheduling, sharing of responsibilities, number and source of entry among others, within each category, "remain flexible for better integration and adaptation to the teachers' individual teaching and assessment styles and practices" (Simon & Forgette-Giroux, 2000, p.89).

Methodology

Eleven volunteer teachers from five school boards in Eastern Ontario, Canada, agreed or asked to apply within their classes, the portfolio assessment content selection framework described above. Five of these teachers taught in two of the three boards that initiated a three-day workshop near the end of the school year to present the framework, while the other six teachers entered the study at various points during the following year. (See Table 1 for a description of the teacher variables). The latter received documentation and coaching on the framework upon request. All teachers were each visited twice from February to May in the year following the three-day workshop. The visits consisted of two in-class observations of portfolio use, followed by a 30 to 45-minute semi-structured interview with each teacher. Of the fifteen general questions, the following four were more or less related to organizational issues surrounding portfolio assessment implementation:

1. How often do you use the portfolio during the week?
2. What responsibilities do the students have toward their portfolios?

3. Was the portfolio used within or across subjects? and
4. Describe any management issues such as storage and format related to the use of the portfolio for assessment purposes.

Observations within the classroom and content analysis of sample portfolios complemented the data obtained from the interview.

Table 1: Input variables for each teacher

	Variables						
Teachers	school board	grade level	discipline	focus of portfolio assessment	year of portfolio use	attendance at a three-day workshop	attendance at group meetings (total of three)
Mark[1]	1	1	Language	Writing	1	-	3
Corey	2	9	Geography	Communication	-	✓*[2]	-*
Gisele	2	7/8	Language	Communication	1	-*	1*
Sally	3	8	All	All	-	✓*	2
John	3	7/8	All	All	-	-*	-
Paul	3	7	Writing	All	-	-*	-
George	4	9	Math/Science	Problem solving	2	-	3*
Frances	4	8	Math	Problem solving	2	-	3*
Edith	5	6	Math	Problem solving	-	-	2*
Joanne	5	7/8	Math	Problem solving	-	-	3*
Carol	5	9	Geography/ Science	Problem solving Critical thinking Team work	-	-	2*

1. Fictional names have been given to protect the anonymity of the participants.
2. An asterisk indicates that the respective school board liaison member also attended

Results

A closer examination of the participants' responses to the questions, their actions, and the content analysis lead to four categories of results. These are presented in the following sections.

Time spent on portfolio assessment

The eleven teachers showed variability in planning and scheduling time around portfolio use. The three teachers from the school board #3 (See Table 1) allowed occasions for the students to select items from a file folder of their best work, to reflect briefly on each item selected, or to mark their own projects. This time was unscheduled and generally seen as extra to their teaching load. The three Grade Nine teachers, on the other hand, planned monthly or biweekly slots of time for their students to select and sort items. The five other teachers, all focusing either on problem solving or communication skills, reserved a full period each Friday or one every day or every other day for portfolio use.

Management issues related to portfolio format, storage and access

Again, all three teachers from school board #3 used a brown accordion style folder with an elastic tie and five inside pockets provided by their board. These were stored in cardboard boxes or in a filing cabinet. The Grade Nine and Grade Eight teachers from school board #2 constructed their portfolios with two large cardboard sheets glued together on three sides. These were stacked haphazardly on a shelf at the back of the classroom. The six others provided their students with manila style folders with two inside pockets all stored in boxes or filing cabinet. All participants had their students keep a "working portfolio" in addition to the assessment portfolio and the students usually did not have free access to their portfolios.

Shared roles

Judging by their comments and actions, George, Paul, John, and Corey tended to underestimate their students' ability to set up, maintain, and reflect on their portfolios. On the few occasions when their students were asked to reflect, they were only required to justify the selection of individual entries. The students' involvement in organizing their portfolios was encouraged primarily to ease teaching tasks. On the other hand, Frances, Edith, and Joanne invited their students to reflect on and self-assess individual applications of problem solving skills. Gisele, Mark and Sally encouraged their students to organize their work themselves and to compare various items within their portfolios using rubrics, checklists, and award stickers.

Context surrounding item selection

The three teachers from school board #3 had their students select entries across subjects but with no clear focus. They had been initially instructed by school board officials to use the portfolio in support of the report card. The three Grade Nine teachers assessed communication or problem solving skills across one or two subjects but felt constrained by the school administrative structure and policies. The Grade One teacher assessed writing skills across Language Arts. In order to holistically assess communications skills across all Language Arts strands using the portfolio, the Grade Eight teacher from school board #2 had her students provide evidence of the framework's five learning dimensions but the entries were not always clearly related to the targeted skill. In stressing problem solving skills, Joanne, Edith, and Frances extended the selection of entries to various disciplines.

Discussion and conclusion

The results suggest the formulation of three sets of research hypotheses. The first alludes to the portfolio assessment implementation process as involving four types of organizational issues: temporal, spatial, human and contextual. Temporal issues concern time spent on planning and scheduling portfolio assessment related activities and their fit within existing teaching and assessment practices. Spatial issues deal with organizing the portfolio's format, physical characteristics, storage, and access. Human aspects include role-sharing such responsibilities as establishing and updating a table of contents, dating and sorting portfolio entries, reflection, and marking for formative or summative assessment purposes. Finally, contextual matters have to do with specifying the object of assessment, determining the scope of disciplines from which portfolio items are selected, and establishing their quantity and quality.

The level of variability among the participating teachers regarding organizational issues suggests a second hypothesis: In implementing portfolio assessment within their classroom, teachers fall along a three or four stage continuum. Novice teachers tend to loosely plan and schedule a rather unfocussed collection of best work across subjects. Storage, access and maintenance are controlled mainly by the teacher. Entries are collected and assessed separately. At the next stage of the continuum, the planned collection over time still remains largely under the responsibility of the teacher but now contains evidence related to the development of a few more or less specified skills or competencies. Students have input in deciding portfolio format, access and storage, and their reflections on and determination of their level of competency are based primarily on the comparison of first drafts to final products within individual assignments. In the final stages, portfolio assessment empowers students to select a minimum number of entries from a variety of contexts in order to provide evidence of the development of all five learning dimensions associated with one or a few clearly articulated competencies. Students regularly reflect on and judge their progress using structured prompts and rubrics that encourage the examination of links and relationships among the portfolio contents.

The data from this study also indicate that particular location and movement of the teachers on the implementation continuum may be a function of variables such as willingness to empower students, previous portfolio experience, school board expectations, training, support and guidance, grade level, and discipline being taught. These factors may be grouped under Myerson's (1997) three generic categories of factors said to affect the implementation process of change within the classroom: teacher uniqueness, professional development, and teaching environment. They also relate to three of Stiggins, & Conklin's (1992) eight assessment environment dimensions: teacher characteristics, teacher perception of students, and policy issues. Whereas the portfolio assessment item selection framework offers specific parameters around assessment purpose, focus, nature, and context, its successful implementation may depend particularly on the extent to which teachers a) accept that portfolio assessment integrates learning and assessment activities, b) obtain training and coaching specifically related to the framework, c) recognize that students are capable and responsible decision-makers with vested interest in self-assessing their own learning, d) learn to better manage the quarter of their professional time they tend to spend on assessment (Stiggins et al., 1992) by planning fewer but complete, sophisticated, and meaningful assessments of competencies involving their students throughout the assessment process, and e) contribute to the development of assessment policies at the school level that facilitate cooperation among teachers, particularly at the high school level. The third set of research hypotheses could focus on the exact nature of the relationships between each of these variables and portfolio assessment implementation in the classroom.

Notes

1. The research reported in this paper was partially supported by a transfer grant from the Ministry of Education and Training, Ontario, Canada to The Ontario Institute for Studies in Education.

2. Details regarding the initial validation study of the framework are reported in Simon, M., & Forgette-Giroux, R. (2000). Impact of a content selection framework on portfolio assessment at the classroom level. *Assessment in Education,* 7(1), 83-101.

References

Calkins, A. (1992). Juneau Portfolio Stories. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Fingeret, H.A. (1993). It Belongs to Me. A Guide to Portfolio Assessment in Adult Education Programs. (ERIC Document Reproduction Service No. ED 359 352)

Glazer, S. M. (1994). User-friendly portfolios: The Search Goes On. *Teaching K-8*, 25 (3), 105-106.

Lescher, M. L. (1998). Portfolio Assessment: Aspects of Implementation and Perspective. An Investigation into the Implementation of Portfolio Assessment in Literacy. Unpublished Doctoral Dissertation, Boston College.

Myerson, M.J. (1997). Naturalistic Assessment: Teacher's Concern and Confidence. (ERIC Document Reproduction Service No. ED394 983).

Sawyer, M. H. (1994). Professional Development and Educational Reform: A Study of Changes in Teachers and Classrooms During Literature Portfolio Implementation. Unpublished doctoral thesis, State University of New York at Albany.

Simon, M., & Forgette-Giroux, R. (2000). Impact of a content selection framework on portfolio assessment at the classroom level. *Assessment in Education*, 7(1), 83-101.

Smit, D., Kolonosky, P, & Seltzer, K. (1991). Implementing a portfolio system. In P. Belanoff, & M. Dickson (Eds.), *Portfolios: Process and Product* (pp. 46-56). Portsmouth, NH: Boynton/Cook Publishers.

Stiggins, R. J., & Conklin, N. F. (1992). *In Teachers's Hands*. Albany: State University of New York Press.

Descriptors: Performance Based Assessments; *Portfolio; Evaluation Problems; Student Evaluation

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Russell, Michael (2000). Summarizing change in test scores: shortcomings of three common methods. *Practical Assessment, Research & Evaluation*, 7(5). Available online: http://ericae.net/pare/getvn.asp?v=7&n=5. This paper has been viewed 3852 times since 5/26/00.

- Find similar papers in
  - ERICAE Full Text Library
  - *Pract Assess, Res & Eval*
  - ERIC RIE & CIJE 1990-
  - ERIC On-Demand Docs
  - ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  - Russell, Michael

Summarizing Change in Test Scores: Shortcomings of Three Common Methods

Michael Russell
Boston College

The reliance on test scores to assess the impact of schools on student achievement has increased sharply during the past decade. This increase is reflected in the number of states that employ testing programs to hold schools, teachers and students accountable for improving student achievement. According to annual surveys by the Council for Chief State School Officers (1998), 48 states use statewide tests to assess student performance in different subject areas and 32 states currently use or plan to use test scores to determine whether to grant diplomas. In addition, many educational programs, including charter schools, depend on test scores to demonstrate the success of their programs. In many cases, however, educational leaders employ overly simplistic and, sometimes, misleading methods to summarize changes in test scores.

Educational leaders, institutions and the popular press have employed a variety of methods to summarize change in test scores. Below, I briefly discuss the advantages and disadvantages of three commonly used methods: Change in Percentile Rank, Scale or Raw Score Change, and Percent Change. A separate article (Russell, 2000) describes two alternate approaches for summarizing change and demonstrates how a third method, namely Expected Growth Size, can be used to summarize change for vertically equated norm-referenced tests.

Method 1: Change in Percentile Rank

As the name implies, the change in percentile rank method focuses on the increase or decrease of the mean percentile ranking for a group of students. Change in test performance is determined by subtracting one mean percentile rank from another.

Since most people are familiar with percentile ranks and the mathematics required for this method are relatively simple, this method is often employed to express change in test scores to the general public. As an example, Edison Schools used this method to report that students, on average, "are gaining more than 5 percentiles per year..." (1999, p. 2).

With this approach, however, two problems can arise. First, calculating the mean percentile rank based on individual percentile ranks can provide an inaccurate estimate of a group's mean performance. Second, due to the unequal intervals separating percentile ranks, changes in mean percentile ranks represent different amounts of growth at each point on the scale.

Why Averaging Ranks to Determine the Mean Group Rank Is Misleading

As Glass and Hopkins (1984) point out, percentile ranks are ordinal measures in which the amount of the trait measured represented by each one point increase in rank varies at each point on the scale. Due to the unequal intervals between each percentile rank, using each individual's percentile rank to determine the mean percentile rank can result in an inaccurate estimate of the group mean. As an example, Table 1 displays the ranks and corresponding times for five sprinters. In this example, the trait measured is running speed. Although the mean rank for the group of sprinters is three, the mean speed for the group is much slower than the time recorded by the third place finisher. Due to the unequal time intervals separating each rank, simply using ranks to determine the mean rank results in an inaccurate estimate of the group's mean running speed.

Table 1: Rank and Finishing Time for Five Sprinters

	Rank	Time
	1	10.2
	2	10.3
	3	10.5
	4	11.2
	5	12.4
Mean	3	10.9

To overcome the problem associated with performing mathematical operations with ordinal ranks, one should use the score associated with each percentile rank to determine the group mean and then determine the percentile rank that corresponds to the group's mean score. As Table 2 demonstrates, using students' standard scores or Normal Curve Equivalent (NCE) scores to calculate the group mean and then finding the percentile rank that corresponds to either mean yields a mean percentile rank that provides a more representative estimate of the group's mean achievement. In Table 2, we see that the mean percentile rank of 27.2 corresponds to a standard score of approximately 163. However, the mean standard score for the group of students is actually 159.3, which corresponds to a percentile rank of approximately 21. In this example, the mean percentile rank over-estimates the group's mean language achievement and implies that on average students are performing six percentiles higher than their mean standard score indicates.

Table 2: Standard Score, NCE and Percentile Rank for Students on the Third Grade ITBS Language Test*

Student	Standard Score	NCE	PR
1	127	1	1
2	143	10	3
3	151	23	10
4	161	35	24
5	164	39	30
6	165	40	32
7	167	42	36
8	169	45	40
9	172	48	46
10	174	50	50
Mean	159.3	33.3	27.2
Corresponding PR	21	21	

*Data based on ITBS Technical Information (Hoover, et al, 1996, p. 33)

Why Differences Between Mean Percentiles Can Be Deceptive

Even when the mean percentile rank is calculated using students' standard scores or NCEs, summarizing change in score as the difference between the group's mean percentile ranks can be misleading because it implies that the same amount of change represents the same amount of growth at all points on the percentile scale. In reality, the further a percentile rank deviates from the mean, the more a student's score must increase for their percentile rank to increase. This relationship is a direct result of the distribution of scores within the normal curve. In a normal curve, a disproportionate number of people score in close proximity to the mean. As a result, a small change in a person's test score close to the mean will result in a much larger change in their rank relative to other test takers as compared to the same change at the extremes of the distribution.

As an example, on the Iowa Test of Basic Skills (ITBS) Language sub-test the standard score for a third-grade student must increase seven points in order to move from the 10th to the 20th percentile. However, to move from the 50th to 60th percentile, a student's standard score only needs to increase four points. Similarly, a ten point increase at the 10th percentile represents a change of about .44 standard deviations. However, a ten-point increase at the 50th percentile represents a change of only .25 standard deviations. Depending upon a student's percentile rank, this method exaggerates or understates change in student performance.

Method 2: Scale or Raw Score Change

A second method used to examine change in test performance focuses on change in scale scores or raw scores. For this method, the mean score for prior years is subtracted from the mean score for the current year. The result represents the change between the two time periods.

As with change in percentile ranks, this method is appealing because it involves basic arithmetic. However, there are several drawbacks. Foremost among them is that when raw scores are used to determine change, it is difficult to compare change across tests that have different score ranges.

As an example, a third-grade mathematics test may contain 30 items while the fourth-grade test contains 40 items. Each grade level may experience a five point increase in its mean score. But, since the tests differ in length, these five point increases do not have the same meaning for both tests. For the test containing 30 items, a five point increase suggests that students are answering about 17% more items correctly. For the test containing 40 items, a five point increase indicates that students are answering only about 13% more items correctly.

One solution to this problem is to focus on change in scale scores. However, this too presents problems. Although most norm-referenced standardized tests report scores from different grade levels on the same scale, the standard deviations for the grade levels tend to differ. For this reason, a five point change for two different grade levels represent different amounts of change within each grade level. The problem is similar to that experienced with the change in percentile rank method. As an example, for the ITBS Language test, the standard deviations for grades 3 and 4 are 19.05 and 24.25, respectively. Thus, five point increases in the standard scores for grade 3 and 4 represents changes of .26 and .21 standard deviations, respectively. Clearly, a five point change represents more growth relative to students within grade 3 than within grade 4.

Method 3: Percent Change

Further distortion is caused by summarizing change in test performance as a percentage of prior performance. As an example, some charter schools have reported 20 to 30% improvements in their test scores. To obtain these figures, test scores for the current year are divided by past test scores to yield the percent change. In the best case scenario, this method focuses on percent change of standard scores or NCEs. In the worst case, this method focuses on percentile ranks. In all cases, however, this method assumes that the scores used to determine percent change are on a ratio measurement scale. Both standard scores and NCEs, however, are at best interval measures while percentile ranks are clearly ordinal measures. As Glass and Hopkins (1984) explain more fully, ratios based on interval and ordinal measures are meaningless.

The percent change method is particularly deceiving when initial performance is low. Take, for example, two schools that both experience five point increases in mean scores. School A saw its mean score increase from 20 to 25, while the mean score for School B increased from 50 to 55. Although both schools experience the same amount of change in their scores, the percent change method suggests that scores for School A improved 25% while School B improved only 10%. Once again, although the arithmetic is simple, the percent change method produces a statistic that is both difficult to interpret and misleading.

Summary

As Willet (1988) explores more fully, all methods of summarizing change may be threatened by low score reliability. However, even when scores are sufficiently reliable, the three methods described above can result in misleading estimates of score changes. In general, these methods are insensitive to the measurement scale on which scores are expressed and perform mathematical operations that are inappropriate for these measurement scales. These methods also assume that the same size difference represents the same amount of change at all points on the scale. As demonstrated above, this assumption is false. For these reasons, all three methods should be avoided when summarizing change in test scores. As is explained more fully in a separate article (see Russell, 2000), preference should be given to methods that yield standardized estimates of score changes which have the same meaning at all points on the measurement scale and which can be compared across tests and grade levels.

References:

Council of Chief State School Officers (1998). *Key State Education Policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance.* Washington, DC.

Edison Schools. (1999). *Second Annual Report on School Performance.* New York, NY.

Glass, G. & Hopkins, K. (1984). *Statistical Methods in Education and Psychology, 2^{nd} Edition*. Boston, MA: Allyn and Bacon.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). *ITBS: Norms and Score Conversions with Technical Information*. Itasca, IL: Riverside Publishing.

Russell, M. (2000). Summarizing change in test scores part II: Advantages of expected growth size estimates. *Practical Assessment, Research and Evaluation*, 7 (6). [Available online: http://ericae.net/pare/getvn.asp?v=7&n=6].

Willett, J. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of Research in Education 15* (pp. 345-422). Washington, DC: American Educational Research Association.

Descriptors: Change; Test Scores; Growth

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Russell, Michael (2000). Using expected growth size estimates to summarize test score changes. *Practical Assessment, Research & Evaluation*, 7(6). Available online: http://ericae.net/pare/getvn.asp?v=7&n=6. This paper has been viewed 2072 times since 5/26/00.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Russell, Michael

Using Expected Growth Size Estimates to Summarize Test Score Changes

Michael Russell
Boston College

An earlier article described the shortcomings of three methods commonly used to summarize changes in test scores (Russell, 2000). This article describes two less commonly used approaches for examining change in test scores, namely Standardized Growth Estimates and Effect Sizes. Aspects of these two approaches are combined and applied to the Iowa Test of Basic Skills (ITBS) to demonstrate the utility of using a third method, termed Expected Growth Size, to examine change in test scores. This article also provides an EXCEL template that readers can use to calculate Expected Growth Size for most standardized tests.

Standardized Growth Estimates

Stenner, Hunter, Bland, & Cooper describe a standardized growth expectation (SGE) as "the amount of growth (expressed in standard deviation form) that a student must demonstrate over a given treatment interval to maintain his/her relative standing in the norm group" (1978, p. 1). To determine an SGE, Stenner et. al. proposed the following three-step method.

> Step 1. The scale score associated with the 50th percentile for a given grade level or the pre-test is identified.

Step 2. The percentile rank for the following grade level or the post-test associated with this scale score is found.

Step 3. The difference between the 50th percentile and the post-test percentile is calculated. To determine this difference, a unit normal deviate table is used to convert percentiles to z-scores and the z-score for the post-test is subtracted from the z-score for the pre-test.

The difference between the pre-test and post-test z-scores is the SGE and expresses "the amount of loss in relative standing that such a student would suffer if he/she learned nothing during the time period" (Stenner, et. al., 1977, p. 1).

As an example, to determine the SGE for grade 3, Table 1 indicates that the scale score associated with the 50th percentile for grade 3 on the ITBS Language sub-test is 174. The percentile rank for grade 4 that corresponds to a scale score of 174 is 26. If a student received the same scale score in grades 3 and 4, their percentile rank would drop from 50 to 26. After both percentiles are converted to z-scores and subtracted, the difference between the two z-scores represents the SGE. In this case, the z-scores corresponding to percentile ranks of 50 and 26 are 0 and -.64, respectively. Thus, the SGE is .64, which indicates a relative loss of .64 standard deviations for a student who shows no change in his/her test score.

Table 1: Percentile Rank, Standard Score and Standard Deviations for the Iowa Test of Basic Skills Language Sub-test

	Percentile Rank	
Standard Score	Grade 3	Grade 4
174	50	26
175	52	27
176	54	29
...	...	...
189	78	47
190	79	48
191	81	50
St. Dev.	19.05	24.25

Effect Sizes

When applying Stenner et. al.'s method for calculating SGEs, Haney, Madaus and Lyons (1993, p. 231-32) point out that the idea of a SGE is analogous to an effect size in that each represents the difference in mean performance of two groups expressed in standard scores. As Glass, McGaw and Smith (1981) describe, an effect size represents the difference between two groups in standard deviations. To calculate an effect size, the difference between the mean of the control group and the experimental group is divided by the standard deviation of the control group. Conceptually, the only difference between an effect size and an SGE is that an effect size is used to compare the means of a "control" group and an "experimental" group while a SGE compares the performance of groups of students at various grade levels.

In the SGE example above, the third grade is designated as the control group and the fourth grade is the experimental group. To determine the effect size or amount of growth between grade three and grade four, the standard score associated with the 50th percentile rank for grade three is subtracted from the standard score associated with the same percentile rank for grade four. This difference is divided by the standard deviation for grade three. Focusing on Table 1, the effect size for grade three is found by subtracting 174 from 191 and dividing by 19.05. The resulting effect size indicates that a student's test score must increase by .89 standard deviations to maintain his/her standing at the 50th percentile.

Expected Growth Size

Although an SGE and an effect size are similar, there is one important difference: an SGE focuses on the standing lost when there is no change in test score, while the effect size focuses on the amount of change in a test score necessary to maintain one's standing. When applied in this manner, the effect size method provides an estimate of the expected growth size between two time periods. In the example above, the expected growth size (EGS) between grade three and grade four on the ITBS Composite Language test is .89 standard deviations.

Defining the Base Year or Control Group

In a well-designed experiment, there is little question as to which group is defined as the control group and which is the experimental group. However, when applying the concept of an effect size to change in test scores between two grade levels, one could reference growth to the pre-test or the post-test distribution.

In the case of SGEs, the post-test distribution is used to reference "growth". Note, however, that although SGEs employ the term growth, the methodology actual provides a measure of loss assuming that a student experiences no growth whatsoever. In this way, using the post-test distribution to reference "growth" is fundamentally flawed in that change is placed in the context of where a student is expected to be rather than from where they started. The situation is analogous to describing someone's progress on trip in relation to how far they still must go in order to reach their destination rather than from how far they have traveled since their departure.

In the case of using an effect size to express growth between two grade levels, one might argue that the pooled standard deviation be employed in lieu of the standard deviation of the control group. However, the difficulty of obtaining an estimate of the pooled standard deviation for most standardized tests forces a choice between designating the pre-test or the post-test as the control group. Given the desire to measure change or growth from where a group begins at one point in time to where they end at a second point in time, the EGS methodology references change to the pre-test distribution. For this reason, the pre-test distribution is assigned as the control group.

Advantages of an Expected Growth Size

Although an expected growth size is more difficult to calculate, it offers three advantages. First, by expressing change in relation to the standard deviation, growth rates for different tests and different grade levels can be compared directly. Table 2 presents expected growth sizes for grades 1 through 8 for several portions of the ITBS. Examining Table 2, one can see that the expected growth sizes differ for each portion of the ITBS. Table 2 also shows an inverse relationship between grade level and size of expected growth. As the grade level increases, the amount of growth students experience decreases.

Table 2: Expected Growth Sizes for the ITBS Reading, Language, Math and Composite Tests

	Growth Size for the 50th Percentile			
Grade Level	Reading	Language	Math	Composite
1	NA	1.46	1.38	1.69
2	.93	1.10	1.25	1.24
3	.79	.89	.89	.99
4	.67	.58	.68	.73
5	.52	.50	.53	.54
6	.39	.32	.42	.43
7	.39	.29	.38	.36
8	.36	.29	.40	.40

Similarly, Table 3 demonstrates that within each grade level, the amount of growth students experience varies by percentile ranks. Students scoring at the 25th percentile experience less growth than students scoring at the mean. And students scoring at the mean experience less growth than students scoring at the 75th percentile. This pattern explains why the standard deviation for most standardized tests increases as the grade level progresses.

Table 3: Expected Growth Sizes for the ITBS Language, Math and Composite Tests By Percentile Rank

	Language			Math			Composite		
	Percentile Rank			Percentile Rank			Percentile Rank		
Grade	25	50	75	25	50	75	25	50	75
1	1.32	1.46	1.68	1.3	1.38	1.5	1.36	1.69	1.65
2	.90	1.1	1.24	1.07	1.25	1.42	1.12	1.24	1.46
3	.60	.89	1.15	.75	.89	1.06	.84	.99	1.2
4	.54	.58	.85	.56	.68	.82	.56	.73	.85
5	.34	.5	.67	.44	.53	.71	.43	.54	.7
6	.23	.32	.42	.35	.42	.52	.36	.43	.51
7	.23	.29	.31	.25	.38	.45	.29	.36	.46
8	.25	.29	.28	.28	.4	.38	.3	.4	.37

Second, once expected growth sizes are calculated for a given test, they can be easily transformed to more common measurement scales. As an example, multiplying the expected growth size by the standard deviation of an Normal Curve Equivalent, NCE, (21.06) provides the number of NCE points a student's score increases during a given time period relative to the student's initial norm group when s/he maintains his/her current standing. For the ITBS Language test, the score for a student who maintains a 50th percentile ranking increases 18.74 NCEs between the third and fourth grade.

Third, once expected growth sizes are transformed to an NCE scale, changes in an individual's or a group's mean score can be reported in relation to expected growth. Performance on most standardized tests is reported relative to the Norm Group for a student's current grade. If the student grows at the same rate as other students in the Norm Group, his/her percentile rank and NCE will remain the same across two years. However, if the student's rate of growth differs from that of the Norm Group, his/her NCE and percentile rank will change.

The expected growth size can be used to determine the extent to which the student's growth exceeded or fell short of the expected growth size. To do so, the student's current NCE is subtracted from his/her previous NCE and divided by the expected NCE growth rate. As an example, consider a student whose NCE for the ITBS Language test increased from 50 in grade 3 to 55 in grade 4. When divided by the expected NCE growth size for third grade (18.74), this five point increase represents 1.27 years of growth. Thus, the student's score increased 27% more than expected.

As Table 2 indicates, growth sizes vary across grade levels. Expressing change in test scores in relation to expected growth size takes these differences in growth rates into consideration. The extent to which performance changes is placed in the context of how scores generally change for students in a given grade. As a result, a more accurate measure of how a student changes relative to other students in his/her grade is produced. As an example, Table 2 shows that students in grade 2 experience about twice as much growth in their test scores compared to students in grade 5. For this reason, an increase of 5 NCEs on the ITBS Composite Math test represents larger growth relative to expected growth for a student in grade 5 than for a student in grade 2.

Limitations of Expected Growth Sizes

Although expected growth sizes provide a sounder approach for summarizing change in test scores than some of the more commonly used approaches, their use is limited to norm referenced standardized tests. Moreover, the EGS methodology assumes that the tests have been vertically equated. When comparing change across multiple years, the methodology also assumes that the tests administered each year provide measures of the same construct based on identical content. Although most norm-referenced tests attempt to meet both assumptions – vertical equating and measures of the same construct – the extent to which they fail to meet these assumptions impacts the accuracy of estimates yielded by the EGS methodology. Finally, as with all comparisons of change over time, the EGS method is also limited by the reliability of the scores used to calculate change. Although there is considerable debate over the extent to which low score reliability impacts the meaningfulness of change scores, caution is advised when employing the EGS method for tests with low reliability (see Willet, 1988 for fuller discussion on reliability and change scores).

Using Expected Growth Sizes for Your Students

To apply expected growth sizes to examine change in the performance of your students, readers are encouraged to use the attached spreadsheet. The spreadsheet provides an easy-to-use template that allows users to calculate expected growth sizes for most standardized tests. In addition, the spreadsheet translates expected growth sizes into expected changes in NCE scores for each grade level.

As the attached instructions indicate, two pieces of information are required to use the spreadsheet: 1. Standard Score to Percentile Rank Conversion tables for the standardized test; and 2. The standard deviation for the standard score for each grade level. This information is available in the Technical Report(s) for each standardized test.

Although expected growth sizes are more complicated to calculate, they provide a more accurate and comparable method of examining change in test scores within and across grade levels and on different tests.

References

Glass, G., McGaw, B. & Smith, M. L. (1981). *Meta-analysis in Social Research.*

Beverly Hills: Sage.

Haney, W., Madaus, G., & Lyons, R. (1993). *The Fractured Marketplace for Standardized Testing*. Boston, MA: Kluwer Academic Publishers.

Russell, M. (2000). Summarizing change in test scores part I: Shortcomings of three common methods. *Practical Assessment, Research and Evaluation*, 7(5). [Available online: http://ericae.net/pare/getvn.asp?v=7&n=5].

Stenner, A. J., Hunter, E. L., Bland, J. D., & Cooper, M. L. (1978). *The standardized growth expectation: Implications for educational evaluation*. Paper presented at the Annual Conference of the American Educational Research Association, Toronto, Canada. (ERIC Document Reproduction Service Number ED169072.)

Willett, J. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of Research in Education 15* (pp. 345-422). Washington, DC: American Educational Research Association.

Descriptors: Change; Test Scores; Growth

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Brem, Sarah, K. & Andrea J. Boyes (2000). Using critical thinking to conduct effective searches of online resources. *Practical Assessment, Research & Evaluation*, 7(7). Available online: http://ericae.net/pare/getvn.asp?v=7&n=7. This paper has been viewed 8069 times since 8/28/00.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

Find articles in ERIC written by
Brem, Sarah, K.
Andrea J. Boyes

Using critical thinking to conduct effective searches of online resources

Sarah K. Brem, Arizona State University
Andrea J. Boyes, Jasper Creek Education, Inc.

While the number of online databases and other resources continues to rise, the quality and effectiveness of database searches does not. Over 80% of academic, public and school libraries offer some form of Internet access (American Library Association, 2000); thousands of full-text electronic journals and serials are available online. However, Hertzberg & Rudner (1999) found that most searches are cursory and ineffective, and they provide extensive recommendations regarding the mechanics of searching. A firm grounding in the mechanics of searching is vital, but an effective search is also an exercise in inquiry and critical thinking. We begin searching a topic with certain questions; as we collect information, we form hypotheses about the topic. These hypotheses in turn guide further searching, and are elaborated, discarded or modified as we learn more.

This document complements guidelines addressing the mechanics of online searching by considering how treating searching as exercises in critical thinking can improve our use of online resources. We address the use of metacognition, hypothesis-testing, and argumentation, providing illustrative examples, and links to tools that can facilitate the process.

METACOGNITION

Metacognition is thinking about thinking (Butler & Wynne, 1995): What do I know? What do I not know? Will I ever find an answer? Knowing what we don't know helps us focus our questions, and how long and hard we look for an answer depends on how likely it seems that we'll find a answer. In the context of online inquiry, it is important to assess how well we're equipped to conduct an inquiry, as well as what's out there to find.

Suppose we want to assess the wisdom of high stakes testing, but are unfamiliar with the issue. We simply enter the phrase "high stakes testing" into ERIC. Doing so retrieves 56 articles. If we quit there, we miss items that would be retrieved by combining terms such as "Accountability-" with "Test-Validity," or "Educational-Testing." These searches would produce an additional 178 articles, enriching our inquiry. At the other end of the spectrum, we may waste time looking for information that no one has, such as how a small subset of the population performs on a particular test. In short, we need to be able to assess the quality of our search.

Once they locate information, people often overlook inconsistencies or conflicts. Searches typically produce a loosely-connected cluster of articles of varying relevance and contrasting opinions. Inquiries are often weakened because disconnected knowledge allows conflicts *between* articles to go undetected; positions are not explicitly compared. In addition, inconsistencies *within* a text may be overlooked because readers tend to form a framework early on--we think we know what the article is about, and miss anything that doesn't fit our framework (Otero & Kintsch, 1992).

How can we improve metacognition in online searching?

Improving metacognition means improving our ability to monitor what we know and how we know it. Here are some ways to accomplish this:

Put the project aside for a brief time. Taking a break helps in several ways. When immersed in the process, people often feel they've learned more than they really have. Nelson and Dunlosky (1991) found that a short break improves the ability to accurately assess what's been learned. Also, returning to a problem repeatedly over time improves memory and comprehension, and allows us to take a slightly different perspective each time.

Talk it out. Chi, deLeeuw, Chiu & LaVancher (1994) find that keeping up a running dialogue with oneself is effective in highlighting inconsistencies and gaps in knowledge. Suppose we read a paper on testing and come across the claim that "passing cutoffs are set arbitrarily." As we attempt to tell ourselves what arbitrary cutoffs means, we realize we don't really know. We can then reread looking for this information, or ferret out additional sources.

Once we've collected a substantial body of knowledge, we can lay out the pros and cons to ourselves or a live audience. Concept mapping can also improve metacognition, and its use is discussed below.

Develop content knowledge. Brem & Rips (in press) found that people who are capable of critical thinking nevertheless fall for weaker arguments when they lack relevant information. Thus, to a certain extent, metacognition and an effective inquiry depend upon building expertise. Nevertheless, we can compensate in the early stages by taking advantage of the content support afforded by online resources.

Many databases provide thesauri--lists of alternative ways of accessing a content area. For example, the ERIC Wizard (http://ericae.net/scripts/ewiz/) uses a thesaurus for widening and narrowing searches. We can construct our own thesauri as well. Examining our initial 56 hits on "high stakes testing," we find other descriptors and keywords associated with these articles--some relevant (Accountability-), some not (Copyrights-); the most relevant become our thesaurus and guide additional searches.

When you don't know, find someone who does. We're often reluctant to admit ignorance, but if we've already tried the strategies above, it's likely that the remaining questions are good, hard questions. Reference librarians, instructors and colleagues can help in locating additional sources and perspectives. Expertise is also available on demand through ERIC Digests and ERIC FAQs (http://ericae.net/nav-lib.htm), which consolidate and synthesize existing information. These documents also help in developing a sense of the overall quality and quantity of evidence available about a topic. For the testing example, ERIC has ten FAQs related to assessment, and nine digests are retrieved by the phrase "high stakes testing." The syntheses of others cannot substitute for working through the issue; in fact, our preparation will help us read these documents with a critical eye and extract relevant information.

HYPOTHESIS TESTING

Searching the literature should be an exercise in hypothesis testing. We hold a certain position on an issue, or construct a position along the way. As we proceed, we need to test and modify this position. The problem is that hypothesis testing is often self-fulfilling. Once we form an opinion, we tend to focus on sources that that support our position, and distort data to make the strongest case (Koehler, 1991). Fortunately, we can combat this process:

How can we improve hypothesis testing in online searching?

Actively pursue alternative hypotheses. We need to fight the tendency to consider only one side of a debate. One of the easiest ways to do this is simply to consider the opposite. Suppose we uncover evidence supporting high-stakes testing. Formulate the opposite opinion--high-stakes testing is a bad idea--and actively work to support this claim. Once we've made an earnest attempt to explore both claims, we can weigh the positions side-by-side.

Develop an evaluativist stance. People frequently fall into an absolutist or multiplist perspective. They see the world in black-and-white, with clear right and wrong answers (Absolutist), or as filled with myriad possibilities, all of which are more or less equally valid (Multiplist). In contrast, adopting an evaluative viewpoint involves recognizing that while there are no right answers, there are better and worse answers, and we can identify them by weighing the evidence. Evaluative approaches are associated with more effective reasoning (Kuhn, 1991), and the strategies described in the next section can aid in the process.

ARGUMENTATION

As we encounter different perspectives, we need a way to decide among them. Which position does the evidence best support? Which sources of evidence and opinions are most reliable? Once we adopt an evaluativist stance, argumentation strategies help us carry out our evaluation.

How can we improve argumentation in online searching?

Consider the structure and reliability of a source. For example, ERIC is a self-contained resource; all information accessed within ERIC meets ERIC standards. In contrast, Web sites often link multiple sources--some more reliable, some less reliable than the site we came from. We need to assess the reliability of every source before we include it in our analysis. Critical thinking guidelines (e.g., Harris, 1997; Kirk, 2000) provide criteria for assessing reliability.

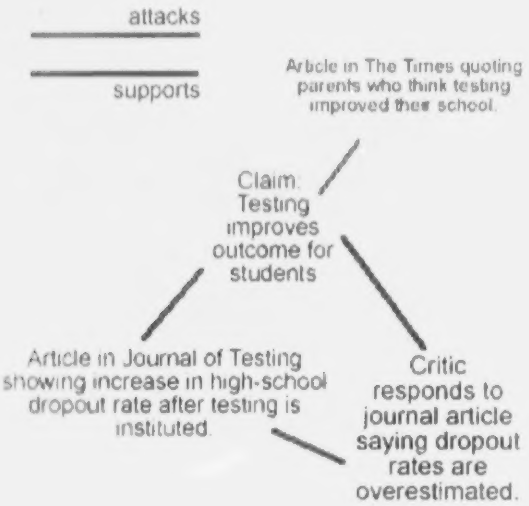
Remember that even reputable sources are fallible. Even the most trusted resource is the work of many people who have different ideas regarding what an article is about and how to describe it. They can make typographical errors. These inconsistencies and mistakes can compromise an inquiry, so it's important to ask whether the results of a search are accurate and complete. The initial goal should be to collect as much relevant information as possible, as it is always possible to narrow the search later.

First, don't initially limit the terms of a search; a broad range of keywords and descriptors increases the likelihood of hitting on the terms chosen by the person entering the data. Second, don't limit which fields are searched. For example, ERIC has "major descriptors" and "minor descriptors;" searching on both maximizes the number of hits. Another example is limiting searches on an author's name to the author field. This seems reasonable, but it misses items with the author's name in the abstract or text; these often present the arguments of opponents and supporters, key pieces of the puzzle. Finally, consider searching on common misspellings, or truncating a term using wildcards to include variations.

Use systematic analysis for a comprehensive (though time-consuming) evaluation. Systematically analyzing an issue takes some time and effort, but generally provides the most complete and accurate evaluation. Systematic analysis involves identifying each claim and asking whether each piece of evidence really supports or refutes it. One popular aid to systematic analysis is using concept maps to visualize the relationship between claims and evidence.

For example, suppose we are searching to see whether we should accept the claim that testing improves student outcomes. We place this claim on the map (Figure 1). When our searches produce a piece of information that supports or attacks this claim, we place a brief description of the evidence on the map and draw lines connecting evidence to claims, choosing lines of different colors or styles to distinguish between supporting and refuting evidence. We also connect pieces of evidence when they attack one another or back each other up. Font size is one way to indicate source reliability (e.g., bigger means more reliable). A map can be made for each alternative viewpoint.

Figure 1. Beginnings of a concept map.
Larger fonts indicate stronger evidence; line color indicates the nature of the relationship.


attacks
supports
Article in The Times quoting parents who think testing improved their school.
Claim: Testing improves outcome for students
Article in Journal of Testing showing increase in high-school dropout rate after testing is instituted.
Critic responds to journal article saying dropout rates are overestimated.

In the resulting visual representation of the debate, a dense web of supporting evidence gives us a solid basis for accepting a claim, and a dense web of refutations provides us with reason to reject. If the evidence seems evenly mixed, or if two alternatives produce equally strong maps, we can continue looking, or we may simply decide that there is no consensus on this issue. In addition, maps support metacognition; holes and smaller text mean holes and weaknesses in the argument, telling us where more information is needed.

Mapping software can facilitate the process (commercial and shareware packages are reviewed at http://www.ozemail.com.au/~caveman/Creative/Software/swindex.htm), but paper and pencil will do. If mapping proves too time-consuming, even a simple list of points for and against a claim is useful. For important decisions, though, mapping is preferred because it includes how claims and evidence are interconnected.

Heuristics are useful when we need to make quick decisions, when there is not enough information for systematic analysis, or to complement systematic approaches. Heuristic evaluation involves making a calculated guess about the quality of an argument. It's usually easy, but not always accurate. For example, deciding to trust someone's argument because they hold a position at a prestigious university is a heuristic--we haven't actually taken the argument apart. It's often a good guess, but even Nobel prize winners have been known to hold a crackpot theory or two. The critical thinking guides mentioned above discuss signs of reliability, and incorporating these into concept maps can enrich our evaluation.

Perhaps the biggest challenge in using heuristics is remembering that a guess is only a guess. This is a metacognitive issue of remembering how we know what we know. Talking out inquiries will help highlight the assumptions underlying heuristics, and using a special color for heuristic contributions to concept maps keeps their status clear.

CONCLUSION

Searching for information online is an exercise in critical thinking, and becoming an expert in critical inquiry takes practice. The guidelines provided above can help in directing and channeling this practice, and providing scaffolding while we gain expertise.

References

American Library Association (2000). LARC Fact Sheet No. 26: How many libraries are on the Internet? [Online] Available: http://www.ala.org/library/fact26.html

Brem, S. K. & Rips, L. J. (in press) Explanation and evidence in informal argument. *Cognitive Science.*

Butler, D., & Winne, P. (1995). Feedback and self-regulated learning: A theoretical synthesis *Review of Educational Research, 65,* 245-281.

Chi, M. T. H., deLeeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477.

Harris, Robert (1997). Evaluating Internet research sources. [Online] Available: http://www.sccu.edu/faculty/R_Harris/evalu8it.htm

Hertzberg, S. & Rudner, L. (1999). The Quality of Researchers' Searches of the ERIC Database. Education Policy Analysis Archives. [Online] Available: http://olam.ed.asu.edu/epaa/v7n25.html

Kirk, E. E. (2000). Evaluating information found on the Internet. [Online] Available: http://milton.mse.jhu.edu:8001/research/education/net.html

Koehler, D. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*, 499-519.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge University Press.

Nelson, T. O. & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The 'delayed-JOL effect.' *Psychological Science,* 2, 267-270.

Otero, J. & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science,* 3, 229-235.

Descriptors: *Critical Thinking; higher order; searching

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

McMillan, James H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation*, 7(8). Available online: http://ericae.net/pare/getvn.asp?v=7&n=8. This paper has been viewed 12703 times since 9/23/00.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  McMillan, James H.

Fundamental Assessment Principles for Teachers and School Administrators

James H. McMillan
Virginia Commonwealth University

While several authors have argued that there are a number of "essential" assessment concepts, principles, techniques, and procedures that teachers and administrators need to know about (e.g. Calfee & Masuda,1997; Cizek, 1997; Ebel, 1962; Farr & Griffin, 1973; Fleming & Chambers, 1983; Gullickson, 1985, 1986; Mayo, 1967; McMillan, 2001; Sanders & Vogel, 1993; Schafer, 1991; Stiggins & Conklin, 1992), there continues to be relatively little emphasis on assessment in the preparation of, or professional development of, teachers and administrators (Stiggins, 2000). In addition to the admonitions of many authors, there are established professional standards for assessment skills of teachers (*Standards for Teacher Competence in Educational Assessment of Students* (1990), a framework of assessment tasks for administrators (Impara & Plake, 1996), the Code of Professional Responsibilities in Educational Measurement (1995), the Code of Fair Testing Practices (1988), and the new edition *of Standards for Educational and Psychological Testing* (1999). If that isn't enough information, a project directed by Arlen Gullickson at The Evaluation Center of Western Michigan University will publish standards for evaluations of students in the near future.

The purpose of this article is to use suggestions and guidelines from these sources, in light of current assessment demands and contemporary theories of learning and motivation, to present eleven "basic principles" to guide the assessment training and professional development of teachers and administrators. That is, what is it about assessment, whether large-scale or classroom, that is fundamental for effective understanding and application? What are the "big ideas" that, when well understood and applied, will effectively guide good assessment practices, regardless of the grade level, subject matter, developer, or user of the results? As Jerome Bruner stated it many years ago in his classic, *The Process of Education*: "...the curriculum of a subject should be determined by the most fundamental understanding that can be achieved of the underlying principles that give structure to that subject." (Bruner, 1960, p.31). What principles, in other words, provide the most essential, fundamental "structure" of assessment knowledge and skills that result in effective educational practices and improved student learning?

Assessment is inherently a process of professional judgment.

The first principle is that professional judgment is the foundation for assessment and, as such, is needed to properly understand and use all aspects of assessment. The measurement of student performance may seem "objective" with such practices as machine scoring and multiple-choice test items, but even these approaches are based on professional assumptions and values. Whether that judgment occurs in constructing test questions, scoring essays, creating rubrics, grading participation, combining scores, or interpreting standardized test scores, the essence of the process is making professional interpretations and decisions. Understanding this principle helps teachers and administrators realize the importance of their own judgments and those of others in evaluating the quality of assessment and the meaning of the results.

Assessment is based on separate but related principles of measurement evidence and evaluation.

It is important to understand the difference between measurement evidence (differentiating degrees of a trait by description or by assigning scores) and evaluation (interpretation of the description or scores). Essential measurement evidence skills include the ability to understand and interpret the meaning of descriptive statistical procedures, including variability, correlation, percentiles, standard scores, growth-scale scores, norming, and principles of combining scores for grading. A conceptual understanding of these techniques is needed (not necessarily knowing how to compute statistics) for such tasks as interpreting student strengths and weaknesses, reliability and validity evidence, grade determination, and making admissions decisions. Schafer (1991) has indicated that these concepts and techniques comprise part of an essential language for educators. They also provide a common basis for communication about "results," interpretation of evidence, and appropriate use of data. This is increasingly important given the pervasiveness of standards-based, high-stakes, large-scale assessments. Evaluation concerns merit and worth of the data as applied to a specific use or context. It involves what Shepard (2000) has described as the systematic analysis of evidence. Like students, teachers and administrators need analysis skills to effectively interpret evidence and make value judgments about the meaning of the results.

Assessment decision-making is influenced by a series of tensions.

Competing purposes, uses, and pressures result in tension for teachers and administrators as they make assessment-related decisions. For example, good teaching is characterized by assessments that motivate and engage students in ways that are consistent with their philosophies of teaching and learning and with theories of development, learning and motivation. Most teachers want to use constructed-response assessments because they believe this kind of testing is best to ascertain student understanding. On the other hand, factors external to the classroom, such as mandated large-scale testing, promote different assessment strategies, such as using selected-response tests and providing practice in objective test-taking (McMillan & Nash, 2000). Further examples of tensions include the following.

- Learning vs auditing
- Formative (informal and ongoing) vs summative (formal and at the end)
- Criterion-referenced vs norm-referenced
- Value-added vs. absolute standards
- Traditional vs alternative
- Authentic vs contrived
- Speeded tests vs power tests
- Standardized tests vs classroom tests

These tensions suggest that decisions about assessment are best made with a full understanding of how different factors influence the nature of the assessment. Once all the alternatives understood, priorities need to be made; trade-offs are inevitable. With an appreciation of the tensions teachers and administrators will hopefully make better informed, better justified assessment decisions.

Assessment influences student motivation and learning.

Grant Wiggins (1998) has used the term 'educative assessment' to describe techniques and issues that educators should consider when they design and use assessments. His message is that the nature of assessment influences what is learned and the degree of meaningful engagement by students in the learning process. While Wiggins contends that assessments should be authentic, with feedback and opportunities for revision to improve rather than simply audit learning, the more general principle is understanding how different assessments affect students. Will students be more engaged if assessment tasks are problem-based? How do students study when they know the test consists of multiple-choice items? What is the nature of feedback, and when is it given to students? How does assessment affect student effort? Answers to such questions help teachers and administrators understand that assessment has powerful effects on motivation and learning. For example, recent research summarized by Black & Wiliam (1998) shows that student self-assessment skills, learned and applied as part of formative assessment, enhances student achievement.

Assessment contains error.

Teachers and administrators need to not only know that there is error in all classroom and standardized assessments, but also more specifically how reliability is determined and how much error is likely. With so much emphasis today on high-stakes testing for promotion, graduation, teacher and administrator accountability, and school accreditation, it is critical that all educators understand concepts like standard error of measurement, reliability coefficients, confidence intervals, and standard setting. Two reliability principles deserve special attention. The first is that reliability refers to scores, not instruments. Second, teachers and administrators need to understand that, typically, error is underestimated. A recent paper by Rogosa (1999), effectively illustrates the concept of underestimation of error by showing in terms of percentile rank probable true score hit-rate and test-retest results.

Good assessment enhances instruction.

Just as assessment impacts student learning and motivation, it also influences the nature of instruction in the classroom. There has been considerable recent literature that has promoted assessment as something that is integrated with instruction, and not an activity that merely audits learning (Shepard, 2000). When assessment is integrated with instruction it informs teachers about what activities and assignments will be most useful, what level of teaching is most appropriate, and how summative assessments provide diagnostic information. For instance, during instruction activities informal, formative assessment helps teachers know when to move on, when to ask more questions, when to give more examples, and what responses to student questions are most appropriate. Standardized test scores, when used appropriately, help teachers understand student strengths and weaknesses to target further instruction.

Good assessment is valid.

Validity is a concept that needs to be fully understood. Like reliability, there are technical terms and issues associated with validity that are essential in helping teachers and administrators make reasonable and appropriate inferences from assessment results (e.g., types of validity evidence, validity generalization, construct underrepresentation, construct-irrelevant variance, and discriminant and convergent evidence). Of critical importance is the concept of evidence based on consequences, a new major validity category in the recently revised *Standards*. Both intended and unintended consequences of assessment need to be examined with appropriate evidence that supports particular arguments or points of view. Of equal importance is getting teachers and administrators to understand their role in gathering and interpreting validity evidence.

Good assessment is fair and ethical.

Arguably, the most important change in the recently published *Standards* is an entire new major section entitled "Fairness in Testing." The *Standards* presents four views of fairness: as absence of bias (e.g., offensiveness and unfair penalization), as equitable treatment, as equality in outcomes, and as opportunity to learn. It includes entire chapters on the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities or special needs. Three additional areas are also important:

- Student knowledge of learning targets and the nature of the assessments prior to instruction (e.g., knowing what will be tested, how it will be graded, scoring criteria, anchors, exemplars, and examples of performance).
- Student prerequisite knowledge and skills, including test-taking skills.
- Avoiding stereotypes.

Good assessments use multiple methods.

Assessment that is fair, leading to valid inferences with a minimum of error, is a series of measures that show student understanding through multiple methods. A complete picture of what students understand and can do is put together in pieces comprised by different approaches to assessment. While testing experts and testing companies stress that important decisions should not be made on the basis of a single test score, some educators at the local level, and some (many?) politicians at the state at the national level, seem determined to violate this principle. There is a need to understand the entire range of assessment techniques and methods, with the realization that each has limitations.

Good assessment is efficient and feasible.

Teachers and school administrators have limited time and resources. Consideration must be given to the efficiency of different approaches to assessment, balancing needs to implement methods required to provide a full understanding with the time needed to develop and implement the methods, and score results. Teacher skills and knowledge are important to consider, as well as the level of support and resources.

Good assessment appropriately incorporates technology.

As technology advances and teachers become more proficient in the use of technology, there will be increased opportunities for teachers and administrators to use computer-based techniques (e.g., item banks, electronic grading, computer-adapted testing, computer-based simulations), Internet resources, and more complex, detailed ways of reporting results. There is, however, a danger that technology will contribute to the mindless use of new resources, such as using items on-line developed by some companies without adequate evidence of reliability, validity, and fairness, and crunching numbers with software programs without sufficient thought about weighting, error, and averaging.

To summarize, what is most essential about assessment is understanding how general, fundamental assessment principles and ideas can be used to enhance student learning and teacher effectiveness. This will be achieved as teachers and administrators learn about conceptual and technical assessment concepts, methods, and procedures, for both large-scale and classroom assessments, and apply these fundamentals to instruction.

Notes:

An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 24, 2000.

References

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan,* 80(2), 139-148.

Bruner, J. S. (1960). *The process of education.* NY: Vintage Books.

Calfee, R. C., & Masuda, W. V. (1997). Classroom assessment as inquiry. In G. D. Phye (Ed.) *Handbook of classroom assessment: Learning, adjustment, and achievement.* NY: Academic Press.

Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.) *Handbook of classroom assessment: Learning, adjustment, and achievement.* NY: Academic Press.

Code of fair testing practices in education (1988). Washington, DC: Joint Committee on Testing Practices (American Psychological Association). Available http://ericae.net/code.htm

Code of professional responsibilities in educational measurement (1995). Washington, DC: National Council on Measurement in Education. Available http://www.unl.edu/buros/article2.html

Ebel, R. L. (1962). Measurement and the teacher. *Educational Leadership*, 20, 20-24.

Farr, R., & Griffin, M. (1973). Measurement gaps in teacher education. *Journal of Research and Development in Education*, 7(1), 19-28.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools*, San Francisco: Jossey-Bass.

Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research, 79*(2), 96-100.

Gullickson, A. R. (1996). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement, 23* (4), 347-354.

Impara, J. C., & Plake, B. S. (1996). Professional development in student assessment for educational administrators. *Educational Measurement: Issues and Practice, 15*(2), 14-19.

Mayo, S. T. (1967). Pre-service preparation of teachers in educational measurement. U.S. Department of Health, Education and Welfare. Washington, DC: Office of Education/Bureau of Research.

McMillan, J. H. (2001). *Essential assessment concepts for teachers and administrators.* Thousand Oaks, CA: Corwin Publishing Company. Available Amazon.com

McMillan, J. H., & Nash, S. (2000). Teachers' classroom assessment and grading decision making. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

Rogosa, D. (1999). How accurate are the STAR national percentile rank scores for individual students? - An interpretive guide. Palo Alto, CA: Stanford University.

Sanders, J. R., & Vogel, S. R. (1993). The development of standards for teacher competence in educational assessment of students, in S. L. Wise (Ed.), *Teacher training in measurement and assessment skills*, Lincoln, NB: Burros Institute of Mental Measurements.

Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice,* 10, (1), 3-6.

Shepard, L. A. (2000). The role of assessment in a learning culture. Paper presented at the Annual Meeting of the American Educational Research Association. Available http://www.aera.net/meeting/am2000/wrap/praddr01.htm

Standards for educational and psychological testing (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Standards for teacher competence in educational assessment of students. (1990). American Federation of Teachers, National Council on Measurement in Education, National Education Association. Available: http://www.unl.edu/buros/article3.html

Stiggins, R. J. (2000). Classroom assessment: A history of neglect, a future of immense potential. Paper presented at the Annual Meeting of the American Educational Research Association.

Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany, NY: State University of New York Press, Albany.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass. Available Amazon.com

Contact Information:

James H. McMillan
Box 842020
Virginia Commonwealth University
Richmond, VA 23284-2020

Phone: 804 828-1332, x553
Fax: 804-225-3554
jmcmilla@saturn.vcu.edu

Descriptors: *Standards; Professional Standards; Test Scores; Student Evaluation

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Alcala, Angelo (2000). A framework for developing an effective instructional program for limited english proficient students with limited formal schooling. *Practical Assessment, Research & Evaluation*, 7(9). Available online: http://ericae.net/pare/getvn.asp?v=7&n=9. This paper has been viewed 3212 times since 9/23/00.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

Find articles in ERIC written by
Alcala, Angelo

A Framework for Developing an Effective Instructional Program for Limited English Proficient Students with Limited Formal Schooling

Angelo Alcala
University of Texas at Arlington

The number of limited English proficient (LEP) children attending American schools has grown dramatically over the past decade. Much of this growth has occurred in states and school districts that previously enrolled only a handful of these students. As the LEP student population has grown, so has the need for the development of special language-learning programs. The challenge of educating LEP students arises from the growth and diversity of this group of students and their diverse academic and social needs (Minicucci & Olsen, 1992).

Until recently, a majority of the secondary schools in the nation were meeting the needs of most of their LEP students by offering courses/programs in English as a Second Language (ESL), which were designed primarily for LEP students possessing literacy skills in their native language. However, with the continuous increases in enrollment of the LEP student population, the number of LEP students lacking literacy skills in their native language has also increased. As a result, there has been an increase in the need for programs designed specifically for this special segment of the LEP student population. This special group of LEP students is most often referred to in the literature as either students with limited formal schooling (LFS) or as "preliterates." Unlike the term "illiterate" which means not knowing how to read and write, the term preliterate implies that the individual will eventually obtain the aforementioned skills.

This article discusses important aspects of the LFS student population:

- LFS student defined,
- impact on schools,
- individualized language development plan,
- classroom instruction, and
- assessment of the LFS student.

WHO IS THE LFS STUDENT?

Generally, the LFS student is an older youth (aged 12-21) who lacks literacy skills in his/her native language because of limited formal education. In most cases, the LFS student possesses less than 2 complete years of a formal education and possesses a language proficiency that is either non-English or limited-English.

Various factors may contribute to the preliterate student's lack of a formal education. The need for the child to supplement the family's income and/or the need for the child to help in the home are two possible factors. Others may include the remote location of a home, the lack of parental supervision, and frequent moves caused by economic need or political turmoil (Morse, 1996; The TESOL Assoc., 1997).

The number of years a student spends in school, the quality of the education received, and the consistency of that educational experience is important in assessing all LEP students. Research indicates that students with strong academic and linguistic skills in the native language will more easily acquire a second language than those with weaker skills (Cummins, 1981). Students who are literate in their native language, who possess grade-level school experience, and who possess an uninterrupted educational background require a very different academic focus than students of the same age who possess only limited, if any, literacy skills in their native language. For instance, a student with limited literacy skills in the native language will require more native-language support than the literate student from the same country. Yet, a majority of the content courses in the typical middle school and high school rely on academic language proficiency in English.

HOW IS THE SCHOOL AFFECTED?

Although the percentage of the LFS youth in the school may only represent a small portion of the LEP student population, the impact can be significant. In most cases, the implementation of additional native language instructional services and the employment or reassignment of instructional assistants to provide these services is necessary. Services provided by these "special" instructional assistants often include instruction, translation between teacher and student, translation between staff and parents, and other language related tasks.

In addition, staff development training for all teachers in topics such as native language instruction, ESL in the content areas, and parental involvement is necessary. Because many of the preliterate students may come from backgrounds very different from those of most teachers, training in multicultural awareness is also important. Teachers must realize that the LFS student population generally finds all aspects of the school experience alien: language, culture, socioeconomic levels, schedules, procedures, and building facilities. Equally as important as the aforementioned topics, although not discussed as often, is the need to train teachers in the utilization of appropriate instructional strategies and the means (authentic assessment) by which to assess LFS students.

HOW DOES ONE DEVELOP AN INDIVIDUALIZED LANGUAGE DEVELOPMENT PLAN (ILDP)?

In an effort to determine the instructional strategies/activities that are most likely to prove effective in working with a particular LFS student, it is highly recommended that an Individualized Language Development Plan (ILDP) for that student be developed. The ILDP, adapted from an individualized education plan developed by Clark and Starr (1996), should serve as the basis for the content, the instructional activities, and the teaching activities that are to be selected for that particular student. In addition, the ILDP should serve as the basis from which to measure the LFS student's progress. The ILDP should include the following:

- an assessment of the student's present level of academic performance (reading/writing in the native language and math)
- an assessment of the student's English language proficiency,
- a diagnosis of the student's strengths and weaknesses,
- a statement of the long-term goals,
- an allocation of the time the student will spend in the selected program (an after school program, a self-contained classroom, a school within a school, a language development center, etc.),
- the person (teacher, parent, specialist, etc.) responsible for each aspect of the instructional service being provided,
- a statement of the short-term instructional goals necessary to attain long-term goals,
- specific recommendations concerning materials of instruction and teaching strategies, and
- appropriate assessment (portfolios, performance, anecdotal records, teacher observations, etc.).

In developing the ILDP, it is also highly recommended that ESL educators take into consideration the ESL Standards for Pre-K-12 Students as set forth by the TESOL Association (1997). The standards are organized by grade-level clusters (pre-K-3, 4-8, 9-12) and address different English proficiency levels (beginning, intermediate, advanced, and limited formal schooling). The purpose of the ESL Standards is to improve the education of students learning English as a second or additional language in the United States.

WHAT TYPES OF INSTRUCTIONAL ACTIVITIES/STRATEGIES BEST MEET THE NEEDS OF THE LFS STUDENT?

First and foremost, prior to considering the instructional activities/strategies to be used, it is extremely important that the student be provided with a warm, caring school/classroom environment. This is vital because, as previously stated, LFS students often find all aspects of the school environment alien. The idea of not "fitting in" can eventually result in the development of low self-esteem and the risk of dropping out (Johnson, Levy, Morales, Morse, and Prokopp, 1986). Past statistics indicate that for many secondary LEP students, the middle school is often the beginning of a high dropout rate (Minicucci, 1985; Olsen & Chenn, 1988).

Varying activities, providing cooperative learning opportunities, and using audio-visual aids while attempting to draw from the student's past experiences is an excellent strategy to implement. The goal is to not only teach students literacy skills in the native language, but to also teach meaningful, communicative, and functional use of the English language. The previously mentioned ESL Standards provide educators with a foundation from which to develop various meaningful opportunities for LFS students to learn English.

For example, giving students an opportunity to communicate (using English) in social settings is Goal 1 Standard 1 of the ESL Standards. According to Holt, Chips, and Wallace (1991), cooperative learning provides the structure for this to occur. In cooperative teams, students with lower levels of proficiency can interact with students who possess a higher level of proficiency in order to negotiate meaning of the content. In this type of learning environment, LFS students can begin to build a strong foundation in oral proficiency as they acquire literacy skills in the second language. Because all students engage in oral practice and utilize interpersonal skills, all students benefit.

According to Goal 1 Descriptors of the ESL Standards, activities like cooperative learning can provide students with an opportunity to share and request information, express needs and feelings, utilize nonverbal communication, engage in conversations, and conduct transactions. Cooperative learning activities can also provide LFS students with the skills that are necessary to function in real-life situations such as the utilization of context for meaning, the seeking of support from others, and the comparing of nonverbal and verbal cues.

Because LFS students are generally older, it is important that school learning result in discourse, products, and performances that have value or meaning in real life beyond success in school. For this reason, some school leaders argue that a distinction be made between academic literacy and functional literacy. Academic literacy is generally identified as that which is free from error in syntax and word structure, punctuation and spelling. Functional literacy, on the other hand, varies according to the individual's needs and divergent roles. These school leaders state that functional literacy rather than academic literacy should be the goal of education for preliterate students (Walker de Felix, Waxman, & Paige, 1994). As a result, many current high school programs have taken this idea a step further and developed courses that provide the LFS student with the training needed to acquire/maintain a job.

Because the focus of well-designed preliterate programs relies heavily on learning that is significant and meaningful in real life, authentic assessment is a must. The goal is to ascertain student progress via a variety of assessment tools. Continuous teacher observations (informal and/or formal), a collection of the student's work samples, and periodic anecdotal descriptions of the student's accomplishments are a few of the methods one can use in assessing the LFS student. To be fully effective, the student and the student's parents should be allowed to participate in assessing whether or not sufficient progress is being made.

Conclusion

Provided that schools recognize and address the special needs of the LFS student population, an LFS student can respond positively with dramatic progress. Although the progress will often vary dramatically from that of the literate LEP student, it is important that teachers recognize it as progress. A proper ILDP, effective instructional strategies, and authentic assessment aid all those involved (the student, the teacher, and the parents) in recognizing the progress made as such. The result is a sense of accomplishment and continued encouragement for learning.

Bibliography

Chamot, A.U. & O'Malley, J.M. (1986). *A cognitive academic language learning approach: An ESL content-based curriculum.* Wheaton: National Clearinghouse for Bilingual Education.

Clark, L., & Starr, I. (1996). *Secondary and middle school teaching methods.* (7th ed.) Upper Saddle, NJ: Prentice Hall.

Collier, V.P. (1992). A synthesis of studies examining long-term language minority student data on academic achievement. *Bilingual Research Journal,* 16, 187-212.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California Department of Education, *Schooling and language minority students: A theoretical framework* (pp. 3-50). Los Angeles, CA: Dissemination and Assessment Center, California State University, Los Angeles.

Hancock, C. (1985). *Teaching pre- and semi-literate Laotian and Cambodian adolescents to read: Helpful hints.* (ERIC Document Reproduction Service No. ED 360 879)

Holt, D.D., Chips, B., & Wallace, D. (1991). *Cooperative learning in the secondary school: Maximizing language acquisition, academic achievement, and social development.* National Clearinghouse for Bilingual Education

Johnson, F., Levy, R., Morales, J., Morse, S., & Prokopp, M. (1986). *Migrant students at the secondary level: Issues and opportunities for change.* Las Cruces, NM: ERIC Clearinghouse on Rural Education and Small Schools. (ED 264 070)

Krashen, S.D. & Terrell, T.D. (2000). *The natural approach.* Pearson Education.

Lado, A. (1990). *Ways in which Spanish-speaking illiterates differ from literates in ESL classrooms.* (ERIC Document Reproduction Service No. ED 367 195)

Minicucci, C. (1985). *Dropping out, losing out: The high cost for California.* Sacramento, CA: Assembly Office of Research.

Minicucci, C., & Olsen, L. (1992). *Programs for secondary limited English proficient students: A California study.* Focus Occasional Paper No. 5 Washington, D.C.: National Clearinghouse for Bilingual Education

Morse, S. (1996). *Unschooled migrant youth: Characteristics and strategies to serve them.* (Temporary Clearing Accession No. RC 020 944)

Olsen, L., & Chen, M. T. (1988). *Crossing the schoolhouse border: Immigrant students and the California public schools.* San Francisco, CA: California Tomorrow.

Richard-Amato, P.A. (1996). *Making it happen: Interaction in the second language classroom.* White Plains, NY: Addison-Wesley.

The TESOL Association. (1997, March). *ESL standards for pre-K-12 students.* Retrieved August 14, 2000 from the World Wide Web: http://www.tesol.org/assoc/k12standards/it/02.html

Walker de Felix, J., Waxman, H.C., & Paige, S. (1994). *Instructional processes in secondary bilingual classrooms.* Third National Research Symposium on Limited English Proficient Student Issues: Focus on Middle and High School Issues.

Descriptors: Limited English proficient; LEP; ESL; English as a Second Language; Limited Formal Schooling; LFS

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2000, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Moskal, Barbara M. & Jon A. Leydens (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10). Available online: http://ericae.net/pare/getvn.asp?v=7&n=10. This paper has been viewed 7908 times since 11/6/00.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Moskal, Barbara M.
  Jon A. Leydens

Scoring Rubric Development: Validity and Reliability

Barbara M. Moskal & Jon A. Leydens
Colorado School of Mines

In Moskal (2000), a framework for developing scoring rubrics was presented and the issues of validity and reliability were given cursory attention. Although many teachers have been exposed to the statistical definitions of the terms "validity" and "reliability" in teacher preparation courses, these courses often do not discuss how these concepts are related to classroom practices (Stiggins, 1999). One purpose of this article is to provide clear definitions of the terms "validity" and "reliability" and illustrate these definitions through examples. A second purpose is to clarify how these issues may be addressed in the development of scoring rubrics. Scoring rubrics are descriptive scoring schemes that are developed by teachers or other evaluators to guide the analysis of the products and/or processes of students' efforts (Brookhart, 1999; Moskal, 2000). The ideas presented here are applicable for anyone using scoring rubrics in the classroom, regardless of the discipline or grade level.

Validity

Validation is the process of accumulating evidence that supports the appropriateness of the inferences that are made of student responses for specified assessment uses. Validity refers to the degree to which the evidence supports that these interpretations are correct and that the manner in which the interpretations are used is appropriate (American Educational Research Association, American

Psychological Association & National Council on Measurement in Education, 1999). Three types of evidence are commonly examined to support the validity of an assessment instrument: content, construct, and criterion. This section begins by defining these types of evidence and is followed by a discussion of how evidence of validity should be considered in the development of scoring rubrics.

Content-Related Evidence

Content-related evidence refers to the extent to which a student's responses to a given assessment instrument reflects that student's knowledge of the content area that is of interest. For example, a history exam in which the questions use complex sentence structures may unintentionally measure students' reading comprehension skills rather than their historical knowledge. A teacher who is interpreting a student's incorrect response may conclude that the student does not have the appropriate historical knowledge when actually that student does not understand the questions. The teacher has misinterpreted the evidence—rendering the interpretation invalid.

Content-related evidence is also concerned with the extent to which the assessment instrument adequately samples the content domain. A mathematics test that primarily includes addition problems would provide inadequate evidence of a student's ability to solve subtraction, multiplication and division problems. Correctly computing fifty addition problems and two multiplication problems does not provide convincing evidence that a student can subtract, multiply or divide.

Content-related evidence should also be considered when developing scoring rubrics. The task shown in Figure 1 was developed by the Quantitative Understanding: Amplifying Student Achievement and Reasoning Project (Lane, et. al, 1995) and requests that the student provide an explanation. The intended content of this task is decimal density. In developing a scoring rubric, a teacher could unintentionally emphasize the nonmathematical components of the task. For example, the resultant scoring criteria may emphasize sentence structure and/or spelling at the expense of the mathematical knowledge that the student displays. The student's score, which is interpreted as an indicator of the student's mathematical knowledge, would actually be a reflection of the student's grammatical skills. Based on this scoring system, the resultant score would be an inaccurate measure of the student's mathematical knowledge. This discussion does not suggest that sentence structure and/or spelling cannot be assessed through this task. If the assessment is intended to examine sentence structure, spelling, *and* mathematics, then the score categories should reflect all of these areas.

Figure 1. Decimal Density Task

Dena tried to identify all the numbers between 3.4 and 3.5. Dena said, "3.41, 3.42, 3.43, 3.44, 3.45, 3.46, 3.47, 3.48 and 3.49. That's all the numbers that are between 3.4 and 3.5."

Nakisha disagreed and said that there were more numbers between 3.4 and 3.5.

A. Which girl is correct?

Answer:

B. Why do you think she is correct?

Construct-Related Evidence

Constructs are processes that are internal to an individual. An example of a construct is an individual's reasoning process. Although reasoning occurs inside a person, it may be partially displayed through results and explanations. An isolated correct answer, however, does not provide clear and convincing evidence of the nature of the individual's underlying reasoning process. Although an answer results from a student's reasoning process, a correct answer may be the outcome of incorrect reasoning. When the purpose of an assessment is to evaluate reasoning, both the product (i.e., the answer) and the process (i.e., the explanation) should be requested and examined.

Consider the problem shown in Figure 1. Part A of this problem requests that the student indicate which girl is correct. Part B requests an explanation. The intention of combining these two questions into a single task is to elicit evidence of the students' reasoning process. If a scoring rubric is used to guide the evaluation of students' responses to this task, then that rubric should contain criteria that addresses both the product and the process. An example of a holistic scoring rubric that examines both the answer and the explanation for this task is shown in Figure 2.

Figure 2. Example Rubric for Decimal Density Task

Proficient:	Answer to part A is Nakisha. Explanation clearly indicates that there are more numbers between the two given values.
Partially Proficient:	Answer to part A is Nakisha. Explanation indicates that there are a finite number of rational numbers between the two given values.
Not	Answer to part A is Dana. Explanation indicates that

Proficient:	all of the values between the two given values are listed.

Note. This rubric is intended as an example and was developed by the authors. It is not the original QUASAR rubric, which employs a five-point scale.

Evaluation criteria within the rubric may also be established that measure factors that are unrelated to the construct of interest. This is similar to the earlier example in which spelling errors were being examined in a mathematics assessment. However, here the concern is whether the elements of the responses being evaluated are appropriate indicators of the underlying construct. If the construct to be examined is reasoning, then spelling errors in the student's explanation are irrelevant to the purpose of the assessment and should not be included in the evaluation criteria. On the other hand, if the purpose of the assessment is to examine spelling and reasoning, then both should be reflected in the evaluation criteria. Construct-related evidence is the evidence that supports that an assessment instrument is completely and only measuring the intended construct.

Reasoning is not the only construct that may be examined through classroom assessments. Problem solving, creativity, writing process, self-esteem, and attitudes are other constructs that a teacher may wish to examine. Regardless of the construct, an effort should be made to identify the facets of the construct that may be displayed and that would provide convincing evidence of the students' underlying processes. These facets should then be carefully considered in the development of the assessment instrument and in the establishment of scoring criteria.

Criterion-Related Evidence

The final type of evidence that will be discussed here is criterion-related evidence. This type of evidence supports the extent to which the results of an assessment correlate with a current or future event. Another way to think of criterion-related evidence is to consider the extent to which the students' performance on the given task may be generalized to other, more relevant activities (Rafilson, 1991).

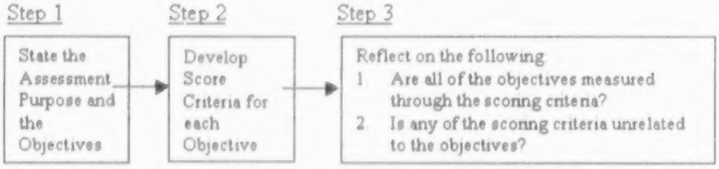
A common practice in many engineering colleges is to develop a course that "mimics" the working environment of a practicing engineer (e.g., Sheppard, & Jeninson, 1997; King, Parker, Grover, Gosink, & Middleton, 1999). These courses are specifically designed to provide the students with experiences in "real" working environments. Evaluations of these courses, which sometimes include the use of scoring rubrics (Leydens & Thompson, 1997; Knecht, Moskal & Pavelich, 2000), are intended to examine how well prepared the students are to function as professional engineers. The quality of the assessment is dependent upon identifying the components of the current environment that will suggest successful performance in the professional environment. When a scoring rubric is used to evaluate performances within these courses, the scoring criteria should address the components of the assessment activity that are directly related to practices in the field. In other words, high scores on the assessment activity should suggest high performance outside the classroom or at the future work place.

Validity Concerns in Rubric Development

Concerns about the valid interpretation of assessment results should begin before the selection or development of a task or an assessment instrument. A well-designed scoring rubric cannot correct for a poorly designed assessment instrument. Since establishing validity is dependent on the purpose of the assessment, teachers should clearly state what they hope to learn about the responding students (i.e., the purpose) and how the students will display these proficiencies (i.e., the objectives). The teacher should use the stated purpose and objectives to guide the development of the scoring rubric.

In order to ensure that an assessment instrument elicits evidence that is appropriate to the desired purpose, Hanny (2000) recommended numbering the intended objectives of a given assessment and then writing the number of the appropriate objective next to the question that addresses that objective. In this manner, any objectives that have not been addressed through the assessment will become apparent. This method for examining an assessment instrument may be modified to evaluate the appropriateness of a scoring rubric. First, clearly state the purpose and objectives of the assessment. Next, develop scoring criteria that address each objective. If one of the objectives is not represented in the score categories, then the rubric is unlikely to provide the evidence necessary to examine the given objective. If some of the scoring criteria are not related to the objectives, then, once again, the appropriateness of the assessment and the rubric is in question. This process for developing a scoring rubric is illustrated in Figure 3.

Figure 3. Evaluating the Appropriateness of Scoring Categories to a Stated Purpose


Step 1
State the Assessment Purpose and the Objectives
Step 2
Develop Score Criteria for each Objective
Step 3
Reflect on the following:
1 Are all of the objectives measured through the scoring criteria?
2 Is any of the scoring criteria unrelated to the objectives?

Reflecting on the purpose and the objectives of the assessment will also suggest which forms of evidence—content, construct, and/or criterion—should be given consideration. If the intention of an assessment instrument is to elicit evidence of an individual's knowledge within a given content area, such as historical facts, then the appropriateness of the content-related evidence should be considered. If the assessment instrument is designed to measure reasoning, problem solving or other processes that are internal to the individual and, therefore, require more indirect examination, then the appropriateness of the construct-related evidence should be examined. If the purpose of the assessment instrument is to elicit evidence of how a student will perform outside of school or in a different situation, criterion-related evidence should be considered.

Being aware of the different types of evidence that support validity throughout the rubric development process is likely to improve the appropriateness of the interpretations when the scoring rubric is used. Validity evidence may also be examined after a preliminary rubric has been established. Table 1 displays a list of questions that may be useful in evaluating the appropriateness of a given scoring rubric with respect to the stated purpose. This table is divided according to the type of evidence being considered.

Table 1: Questions to Examine Each Type of Validity Evidence

Content	Construct	Criterion
1. Do the evaluation criteria address any extraneous content? 2. Do the evaluation criteria of the scoring rubric address all aspects of the intended content? 3. Is there any content addressed in the task that should be evaluated through the rubric, but is not?	1. Are all of the important facets of the intended construct evaluated through the scoring criteria? 2. Is any of the evaluation criteria irrelevant to the construct of interest?	1. How do the scoring criteria reflect competencies that would suggest success on future or related performances? 2. What are the important components of the future or related performance that may be evaluated through the use of the assessment instrument? 3. How do the scoring criteria measure the important components of the future or related performance? 4. Are there any facets of the future or related performance that are not reflected in the scoring criteria?

Many assessments serve multiple purposes. For example, the problem displayed in Figure 1 was designed to measure both students' knowledge of decimal density and the reasoning process that students used to solve the problem. When multiple purposes are served by a given assessment, more than one form of evidence may need to be considered.

Another form of validity evidence that is often discussed is "consequential evidence". Consequential evidence refers to examining the consequences or uses of the assessment results. For example, a teacher may find that the application of the scoring rubric to the evaluation of male and female performances on a given task consistently results in lower evaluations for the male students. The interpretation of this result may be the male students are not as proficient within the area that is being investigated as the female students. It is possible that the identified difference is actually the result of a factor that is unrelated to the purpose of the assessment. In other words, the completion of the task may require knowledge of content or constructs that were not consistent with the original purposes. Consequential evidence refers to examining the outcomes of an assessment and using these outcomes to identify possible alternative interpretations of the assessment results (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999).

Reliability

Reliability refers to the consistency of assessment scores. For example, on a reliable test, a student would expect to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response. On an unreliable examination, a student's score may vary based on factors that are not related to the purpose of the assessment.

Many teachers are probably familiar with the terms "test/retest reliability," "equivalent-forms reliability," "split half reliability" and "rational equivalence reliability" (Gay, 1987). Each of these terms refers to statistical methods that are used to establish consistency of student performances within a given test or across more than one test. These types of reliability are of more concern on standardized or high stakes testing than they are in classroom assessment. In a classroom, students' knowledge is repeatedly assessed and this allows the teacher to adjust as new insights are acquired.

The two forms of reliability that typically are considered in classroom assessment and in rubric development involve rater (or scorer) reliability. Rater reliability generally refers to the consistency of scores that are assigned by two independent raters and that are assigned by the same rater at different points in time. The former is referred to as "interrater reliability" while the latter is referred to as "intrarater reliability."

Interrater Reliability

Interrater reliability refers to the concern that a student's score may vary from rater to rater. Students often criticize exams in which their score appears to be based on the subjective judgment of their instructor. For example, one manner in which to analyze an essay exam is to read through the students' responses and make judgments as to the quality of the students' written products. Without set criteria to guide the rating process, two independent raters may not assign the same score to a given response. Each rater has his or her own evaluation criteria. Scoring rubrics respond to this concern by formalizing the criteria at each score level. The descriptions of the score levels are used to guide the evaluation process. Although scoring rubrics do not completely eliminate variations between raters, a well-designed scoring rubric can reduce the occurrence of these discrepancies.

Intrarater Reliability

Factors that are external to the purpose of the assessment can impact the manner in which a given rater scores student responses. For example, a rater may become fatigued with the scoring process and devote less attention to the analysis over time. Certain responses may receive different scores than they would have had they been scored earlier in the evaluation. A rater's mood on the given day or knowing who a respondent is may also impact the scoring process. A correct response from a failing student may be more critically analyzed than an identical response from a student who is known to perform well. Intrarater reliability refers to each of these situations in which the scoring process of a given rater changes over time. The inconsistencies in the scoring process result from influences that are internal to the rater rather than true differences in student performances. Well-designed scoring rubrics respond to the concern of intrarater reliability by establishing a description of the scoring criteria in advance. Throughout the scoring process, the rater should revisit the established criteria in order to ensure that consistency is maintained.

Reliability Concerns in Rubric Development

Clarifying the scoring rubric is likely to improve both interrater and intrarater reliability. A scoring rubric with well-defined score categories should assist in maintaining consistent scoring regardless of who the rater is or when the rating is completed. The following questions may be used to evaluate the clarity of a given rubric: 1) Are the scoring categories well defined? 2) Are the differences between the score categories clear? And 3) Would two independent raters arrive at the same score for a given response based on the scoring rubric? If the answer to any of these questions is "no", then the unclear score categories should be revised.

One method of further clarifying a scoring rubric is through the use of anchor papers. Anchor papers are a set of scored responses that illustrate the nuances of the scoring rubric. A given rater may refer to the anchor papers throughout the scoring process to illuminate the differences between the score levels.

After every effort has been made to clarify the scoring categories, other teachers may be asked to use the rubric and the anchor papers to evaluate a sample set of responses. Any discrepancies between the scores that are assigned by the teachers will suggest which components of the scoring rubric require further explanation. Any differences in interpretation should be discussed and appropriate adjustments to the scoring rubric should be negotiated. Although this negotiation process can be time consuming, it can also greatly enhance reliability (Yancey, 1999).

Another reliability concern is the appropriateness of the given scoring rubric to the population of responding students. A scoring rubric that consistently measures the performances of one set of students may not consistently measure the performances of a different set of students. For example, if a task is embedded within a context, one population of students may be familiar with that context and the other population may be unfamiliar with that context. The students who are unfamiliar with the given context may achieve a lower score based on their lack of knowledge of the context. If these same students had completed a different task that covered the same material that was embedded in a familiar context, their scores may have been higher. When the cause of variation in performance and the resulting scores is unrelated to the purpose of the assessment, the scores are unreliable.

Sometimes during the scoring process, teachers realize that they hold implicit criteria that are not stated in the scoring rubric. Whenever possible, the scoring rubric should be shared with the students in advance in order to allow students the opportunity to construct the response with the intention of providing convincing evidence that they have met the criteria. If the scoring rubric is shared with the students prior to the evaluation, students should not be held accountable for the unstated criteria. Identifying implicit criteria can help the teacher refine the scoring rubric for future assessments.

Concluding Remarks

Establishing reliability is a prerequisite for establishing validity (Gay, 1987). Although a valid assessment is by necessity reliable, the contrary is not true. A reliable assessment is not necessarily valid. A scoring rubric is likely to result in invalid interpretations, for example, when the scoring criteria are focused on an element of the response that is not related to the purpose of the assessment. The score criteria may be so well stated that any given response would receive the same score regardless of who the rater is or when the response is scored.

A final word of caution is necessary concerning the development of scoring rubrics. Scoring rubrics describe general, synthesized criteria that are witnessed across individual performances and therefore, cannot possibly account for the unique characteristics of every performance (Delandshere & Petrosky, 1998; Haswell & Wyche-Smith, 1994). Teachers who depend solely upon the scoring criteria during the evaluation process may be less likely to recognize inconsistencies that emerge between the observed performances and the resultant score. For example, a reliable scoring rubric may be developed and used to evaluate the performances of pre-service teachers while those individuals are providing instruction. The existence of scoring criteria may shift the rater's focus from the *interpretation* of an individual teacher's performances to the mere *recognition* of traits that appear on the rubric (Delandshere & Petrosky, 1998). A pre-service teacher who has a unique, but effective style, may acquire an invalid, low score based on the traits of the performance.

The purpose of this article was to define the concepts of validity and reliability and to explain how these concepts are related to scoring rubric development. The reader may have noticed that the different types of scoring rubrics—analytic, holistic, task specific, and general—were not discussed here (for more on these, see Moskal, 2000). Neither validity nor reliability is dependent upon the type of rubric. Carefully designed analytic, holistic, task specific, and general scoring rubrics have the potential to produce valid and reliable results.

REFERENCES

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Delandshere, G. & Petrosky, A. (1998) "Assessment of complex performances: Limitations of key measurement assumptions." *Educational Researcher, 27* (2), 14-25.

Gay, L.R. (1987). "Selection of measurement instruments." In *Educational Research: Competencies for Analysis and Application* (3rd ed.). New York: Macmillan.

Hanny, R. J. (2000). *Assessing the SOL in classrooms.* College of William and Mary. [Available online: http://www.wm.edu/education/SURN/solass.html].

Haswell, R., & Wyche-Smith, S. (1994) "Adventuring into writing assessment." *College Composition and Communication, 45,* 220-236.

King, R.H., Parker, T.E., Grover, T.P., Gosink, J.P. & Middleton, N.T. (1999). "A multidisciplinary engineering laboratory course." *Journal of Engineering Education, 88* (3) 311- 316.

Knecht, R., Moskal, B. & Pavelich, M. (2000). *The design report rubric: Measuring and tracking growth through success.* Proceedings of the Annual Meeting American Society for Engineering Education, St. Louis, Missouri.

Lane, S., Silver, E.A., Ankenmann, R.D., Cai, J., Finseth, C., Liu, M., Magone, M.E., Meel, D., Moskal, B., Parke, C.S., Stone, C.A., Wang, N.,

& Zhu, Y. (1995). *QUASAR Cognitive Assessment Instrument (QCAI).* Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Leydens, J. & Thompson, D. (1997, August). *Writing rubrics design (EPICS) I,* Internal Communication, Design (EPICS) Program, Colorado School of Mines.

Moskal, B. M. (2000). "Scoring rubrics: What, when and how?" *Practical Assessment, Research & Evaluation, 7* (3) [Available Online: http://ericae.net/pare/getvn.asp?v=7&n=3].

Rafilson, F. (1991). "The case for validity generalization." *Practical Assessment, Research & Evaluation, 2* (13). [Available online: http://ericae.net/pare/getvn.asp?v=2&n=13].

Sheppard, S. & Jeninson, R. (1997). "Freshman engineering design experiences and organizational framework." *International Journal of Engineering Education, 13* (3), 190-197.

Stiggins, R. J. (1999). "Evaluating classroom assessment training in teacher education programs." *Educational Measurement: Issues and Practice, 18* (1), 23-27.

Yancey, K.B. (1999). "Looking back as we look forward: Historicizing writing assessment." *College Composition and Communication, 50,* 483-503.

Authors

Barbara M. Moskal
Associate Director of the Center for Engineering Education
Assistant Professor of Mathematical and Computer Sciences
Colorado School of Mines
1500 Illinois St.
Golden, Colorado 80401

Jon A. Leydens
Writing Program Administrator
Division of Liberal Arts and International Studies
Colorado School of Mines
1500 Illinois St.
Golden, Colorado 80401

Descriptors: *Rubrics; Scoring; *Student Evaluation; *Test Construction; *Evaluation Methods; Grades; Grading; *Scoring; Reliability; Validity

Practical Assessmen Research & Evaluation

Volume 7, 2001 **Articles 11-22**

Article Pages

11 *Robert L. Linn*: Assessments and Accountability (Condensed Version) 5

12 *Jerrell C. Cassady:* Self-Reported GPA and SAT: A Methodological Note 7

13 *Christopher Tienken & Michael Wilson*: Using State Standards and Test to Improve Instruction 9

14 *Lawrence M. Rudner*: Computing the Expected Proportions of Misclassified Examinees 7

15 *William D. Schafer*: Replication in Field Research 8

16 *Cody S. Ding*: Profile Analysis: Multidimensional Scaling Approach 10

17 *Steve Stemler*: An Overview of Content Analysis 10

18 *Marielle Simon & Renée Forgette-Giroux*: A Rubric for Scoring Postsecondary Academic Skills 7

19 *David J. Solomon*: Conducting Web-Based Surveys 6

20 *Jerrell C. Cassady*: The Stability of Undergraduate Students' Cognitive Test Anxiety Levels 8

21 *Paul M. La Marca*: Alignment of Standards and Assessments as an Accountability Criterion 7

22 *Gordon E. Taub*: A Confirmatory Analysis of the Wechsler Adult Intelligence Scale-Third Edition: Is the Verbal/Performance Discrepancy Justified? 9

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ [ ] Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Linn, Robert L. (2001). Assessments and accountability (condensed version). *Practical Assessment, Research & Evaluation*, 7(11). Available online: http://ericae.net/pare/getvn.asp?v=7&n=11. This paper has been viewed 5023 times since 1/16/01.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Linn, Robert L.

Assessments and Accountability (Condensed version)

Robert L. Linn
Center for Research on Evaluation, Standards, and Student Testing

Adapted, with permission of Robert L. Linn and the American Educational Research Association, from Linn, R. L. (2000). Assessments and accountability. Educational Researcher, 29 (2), 4-16.

Assessment and accountability have played prominent roles in many of the education reform efforts during the past 50 years. In the 1950s, under the influence of James B. Conant's work on comprehensive high schools, testing was used to select students for higher education and to identify students for gifted programs. By the mid-1960s test results were used as one measure to evaluate the effectiveness of Title I and other federal programs. In the 1970s and early 1980s, the minimum competency testing movement spread rapidly; 34 states instituted some sort of testing of basic skills as a graduation requirement. Overlapping the minimum competency testing movement and continuing into the late 1980s and early 1990s was the expansion of the use of standardized test results for accountability purposes.

Assessment is appealing to policymakers for several reasons: it is relatively inexpensive compared to making program changes, it can be externally mandated, it can be implemented rapidly, and it offers visible results. This Digest discusses significant features of present-day assessment programs and offers recommendations to increase positive effects and minimize negative ones.

What Are the Characteristics of Current Reform Efforts?

Although a number of other important features might be considered in any discussion of assessment and education reform (e.g., the emphasis on performance-based approaches to assessment, the concept of tests worth teaching to, and the politically controversial and technically challenging issue of opportunity to learn), I focus on the following three:

- An emphasis on the development and use of ambitious content standards as the basis of assessment and accountability.
- The dual emphasis on setting demanding performance standards and on the inclusion of all students.
- The attachment of high-stakes accountability mechanisms for schools, teachers, and sometimes, students.

Content standards. The federal government has encouraged states to develop content and performance standards that are demanding. Standards-based reform is also a central part of many of the state reform efforts, including ones such as Kentucky and Maryland that have been using standards-based assessments for several years and ones such as Colorado and Missouri that have more recently introduced standards-based assessment systems. A great deal has been written about the strengths and weaknesses of content standards (e.g., *Education Week*, 1997; Lerner, 1998; Olson, 1998; Raimi & Braden, 1998).

It is worth acknowledging that content standards vary a good deal in specificity and in emphasis. Content standards can, and should, if they are to be more than window dressing, influence both the choice of constructs to be measured and the ways in which they are eventually measured.

Performance standards. Performance standards are supposed to specify how good is good enough. There are at least four critical characteristics of performance standards. First, they are intended to be absolute rather than normative. Second, they are expected to be set at high, world-class levels. Third, a relatively small number of levels (e.g., advanced, proficient) are typically identified. Finally, they are expected to apply to all, or essentially all, students, rather than a selected subset such as college-bound students seeking advanced placement.

Should the intent be to aspire not just to high standards for all students, but to the same high standards for all students and on the same time schedule for all students (e.g., meet reading standards in English at the end of Grade 4)? Coffman (1993) sums up the problems of holding common high standards for all students as follows: "Holding common standards for all pupils can only encourage a narrowing of educational experiences for most pupils, doom many to failure, and limit the development of many worthy talents" (p. 8). Although this statement runs counter to the current zeitgeist and may not even be considered politically correct, it seems to me a sensible conclusion that is consistent with both evidence and common sense. Having high standards is not the same as having common standards for all, especially when they are tied to a lock step of age or grade level.

High-stakes accountability. The use of student performance on tests in accountability systems is not new. Examples of payment for results such as the flurry of performance contracting in the 1960s can be found cropping up and fading away over many decades. What is somewhat different about the current emphasis on performance-based accountability is its pervasiveness. As Elmore, Abelmann, and Fuhrman note, "What is new is an increasing emphasis on student performance as the touchstone for state governance" (1996, p. 65). Student achievement is being used not only to single out schools that require special assistance, but also to provide cash incentives for imrovements in performance. Yet several fundamental questions remain about the student assessments, the accountability model, and the validity, impact, and credibility of the system.

As noted earlier, for example, the choice of constructs matters. Content areas (and subareas within those content areas) that are assessed for a high-stakes accountability receive emphasis while those that are left out languish. Meyer (1996) has argued that "in a high-stakes accountability system, teachers and administrators are likely to exploit all avenues to improve measured performance. For example, teachers may 'teach narrowly to the test.' For tests that are relatively immune to this type of corruption, teaching to the test could induce teachers and administrators to adopt new curriculums and teaching techniques much more rapidly than they otherwise would" (p. 140).

It is unclear, however, that there is either the know-how or the will to develop assessments that are sufficiently "immune to this type of corruption." It is expensive to introduce a new, albeit well-equated, form of a test on each new administration. And if ambitious performance-based tasks are added to the mix, still greater increases in costs will result.

A second area of concern regarding high-stakes assessments relates to what data the basic model should employ. Some possibilities include current status, comparisons of cross-sectional cohorts of students at different grades in the same year, comparisons of cross-sectional cohorts in a fixed grade from one year to the next, longitudinal comparisons of school aggregate scores without requiring matched individual data, and longitudinal comparisons based only on matched student records. Should simple change scores be used or some form of regression-based adjustment? And, if regression-based adjustments are used, what variables should be included as predictors? In particular, should measures of socioeconomic status be used in the adjustments?

Elmore, Abelmann, and Furhman (1996) present both sides of this issue, noting that on the one hand, schools can fairly be held accountable only for those factors they can control, but on the other, controlling for student background or prior achievement institutionalizes low expectations for poor, minority, low-achieving students (pp. 93-94). Kentucky's interesting approach to this dilemma has been to set a common goal for all schools by the end of 20 years, thus establishing faster biennial growth targets for initially low-scoring schools than initially high-scoring schools (Guskey, 1994).

The biggest question of all is whether the assessment-based accountability models that are now being used or being considered by states and districts have been shown to improve education. Unfortunately, it is difficult to get a clear-cut answer to this simple question. Certainly, there is evidence that performance on the measures used in accountability systems increases over time, but that can also be linked to the use of old norms, the repeated use of test forms year after year, the exclusion of students from participating in accountability testing programs, and the narrow focusing of instruction on the skills and question types used on the test (see

Koretz, 1988; Linn et al., 1990; Shepard, 1990). Comparative data are needed to evaluate the apparent gains. The National Assessment of Educational Progress provides one source of such data. Comparisons of state NAEP and state assessment results sometimes suggest similar trends; for example, increases in numbers of students scoring at or above basic or proficient levels on NAEP may track with improved state test scores over time. In other cases, the trends for a state's own assessment and NAEP will suggest contradictory conclusions about the changes in student achievement. Divergence of trends does not prove that NAEP is right and the state assessment is misleading, but it does raise important questions about the generalizability of gains reported on a state's own assessment, and hence, about the validity of claims regarding student achievement.

How Can Assessments Be Used More Wisely?

Assessment systems that are useful monitors lose much of their dependability and credibility for that purpose when high stakes are attached to them. The unintended negative effects of the high-stakes accountability uses often outweigh the intended positive effects. It is worth arguing for more modest claims about uses that can validly be made of our best assessments and warning against the over-reliance on them that is so prevalent and popular. To enhance the validity, credibility, and positive impact of assessment and accountability systems while minimizing their negative effects, policymakers should:

1. Provide safeguards against selective exclusion of students from assessments.
2. Make the case that high-stakes accountability requires new high-quality assessments each year that are equated to those of previous years.
3. Don't put all of the weight on a single test. Instead, seek multiple indicators. The choice of construct matters and the use of multiple indicators increases the validity of inferences based upon observed gains in achievement.
4. Place more emphasis on comparisons of performance from year to year than from school to school. This allows for differences in starting points while maintaining an expectation of improvement for all.
5. Consider both value added and status in the system. Value added provides schools that start out far from the mark a reasonable chance to show improvement while status guards against institutionalizing low expectations for those same students and schools.
6. Recognize, evaluate, and report the degree of uncertainty in the reported results.
7. Put in place a system for evaluating both the intended positive effects and the more likely unintended negative effects of the system.

References

Coffman, W. E. (1993). A king over Egypt, which knew not Joseph. *Educational Measurement: Issues and Practice*, 12(2), 5-8.

Education Week (1997, January 22). Quality counts: A report card on the condition of public education in the 50 states. *A Supplement to Education Week*, vol. 16.

Elmore, R. F., Abelmann, C. H., & Fuhrman, S. H. (1996). The new accountability in state education reform: From process to performance. In H. F. Ladd (Ed..), *Holding schools accountable: Performance-based reform in education* (pp. 65-98). Washington, DC: The Brookings Institution.

Guskey, T. R. (Ed.) (1994). *High stakes performance assessment: Perspectives on Kentucky's reform*. Thousand Oaks, CA: Corwin Press.

Koretz, D. (1988). Arriving at Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator 12*(2), 8-15, 46-52.

Lerner, L. S. (1998). *State science standards: An appraisal of science standards in 36 states.* Washington, D.C. Thomas B. Fordham Foundation.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice 9*(3), 5-14.

Meyer, R. H. (1996). Comments on chapters two, three, and four. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 137-145). Washington, DC: The Brookings Institution.

Olson, L. (1998, April 15). An "A" or a "D": State rankings differ widely. *Education Week 17*, 1, 18.

Raimi, R. A., & Braden, L. S. (1998). State mathematics standards: An appraisal of science standards in 46 states, the District of Columbia, and Japan. Washington, DC.: The Thomas B. Fordham Foundation.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, *9*(3), 15-22.

Descriptors: Academic Achievement; Academic Standards; * Accountability; Educational Change; Educational History; Elementary Secondary Education; * Minimum Competency Testing; Standardized Tests; * Student Evaluation; Reform Efforts

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Cassady, Jerrell C. (2001). Self-reported gpa and sat: a methodological note. *Practical Assessment, Research & Evaluation*, 7(12). Available online: http://ericae.net/pare/getvn.asp?v=7&n=12. This paper has been viewed 1156 times since 1/24/01.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Cassady, Jerrell C.

Self-Reported GPA and SAT: A Methodological Note

Jerrell C. Cassady
Ball State University

The use of self-reports from students is a common, yet risky, methodological venture in psychological research. Relying upon an individual to provide an accurate and unbiased rating of her or his ability, attitude, or past experiences is problematic, considering that the behavioral revolution in psychology was driven in part by a rejection of the reliance on malleable data sources. Researchers now concentrate on observable phenomena that can be validated; however, they are often forced to ask participants to provide various forms of data without having the means to verify their accuracy.

The research base demonstrating the magnitude of disparity between actual and reported performance scores is very limited. This article investigates the methodological practice of relying on self-reported Scholastic Assessment Test (SAT) and college grade point average (GPA) scores provided by undergraduate research participants. It also attempts to provide an explanation for the differential reliability of self-reported SAT and GPA values by examining the differences in students' access to these two scores. Because most undergraduates see their GPAs every 16 to 20 weeks, but seldom review their SAT scores after college entrance, it is expected that reliability will be greater for self-reported GPA scores than for self-reported SAT scores among undergraduates.

The research in SAT score accuracy has generally indicated that students' reports correlate with actual scores in the range of .60 to .80 (Goldman, Flake, and Matheson, 1990; Frucot and

Cook, 1994; Trice, 1990). Furthermore, there is evidence that individuals who do not provide their scores are more likely to have low SAT scores, suggesting a potential skew in the self-report performance literature (Flake and Goldman, 1991; Trice, 1990). In a rigorous analysis of the relationship between actual and reported scores on the SAT, Shepperd (1993) reported that students with low SAT scores not only inflated their self-reported scores, but also rated the score they received on the SAT as inaccurate or flawed. Furthermore, when students reported SAT scores with no explicit instructions, the tendency to inflate the score was evident. However, when the students were asked to report their SAT scores for a second time (two months after the initial report), but with an incentive for accuracy and the assurance that any inflation would be known, the average deviation from true score was 9 points for the total scale SAT score, a mere 1/10 of a standard deviation (Shepperd, 1993). Shepperd hypothesized that this pattern supported the theory that the inflation was an attempt to portray a positive image, rather than a misrepresentation arising from a memory deficit. As for GPA ratings, there is also evidence for skewed self-reports; specifically, there is greater inflation by students with lower GPAs than by students with higher GPAs (Dobbins, Farh, and Werbel, 1993; Frucot and Cook, 1994). This inflation of GPA has been found to be free from a ceiling effect, and it has been proposed to be a consequence of social desirability (Dobbins et al., 1993).

Altogether, there are insufficient data regarding the validity of self-reported SAT and GPA values to be confident in this common methodological practice. Furthermore, the existing reports often do not provide an indication of the absolute, or relative, degree of deviation. This study was conducted to test the accuracy and trends of deviation noted in undergraduates' self-reported SAT and GPA values. The magnitude of deviation was examined through two independent variables, the direction of deviation (if any) and the actual performance level of the student on the measure in question. The results were expected to support previous reports that self-reported values for GPA and SAT were relatively reliable (in the range of .70 to .90). Furthermore, the results were expected to identify that for both GPA and SAT, low scorers' ratings would vary from actual scores more than high scorers, with the self-reported values demonstrating an inflated value. Finally, it was predicted that individuals who overestimated their performance levels would do so to a greater magnitude than those individuals reporting an underestimation of their performance.

Method

Participants

Eighty-nine undergraduate students at a mid-sized Midwestern university reported their current cumulative GPA and the scores they had received on the SAT. Ninety-six percent of the participants (n = 86) identified themselves as Caucasian; the remaining three students identified themselves as African American. Eighty-nine percent (n = 79) were female. All participants were in the second year of the undergraduate preservice teacher education program. Participants reported ages ranging from 19 to 28 (M = 19.99, SD = 1.06).

Procedure

The participants were asked to provide their undergraduate cumulative GPAs and their official SAT scores as part of another research project. Those participants who indicated they were unsure of their scores were asked to provide their "best guess" regarding the SAT verbal, math, and total scores, as well as GPA. There was no indication to the students at the time they provided their scores what the scores would be used for, or that the scores would be checked against their official records. The participants were debriefed in a subsequent experimental session, at which time they provided consent to access the necessary university records.

Students who did not take the SAT (typically taking the ACT for entrance) were excluded from the analyses of SAT score accuracy. Similarly, students without official university grade records (i.e., transfer students from community colleges) were excluded from the analyses on the accuracy of GPA. After the participants granted consent to access the records, their actual SAT and GPA values were gathered from official university records during the same semester to ensure that the students' official GPA was not affected by grades not finalized at the time of data collection.

Analyses

Initial investigation of the data relied on examining the correlation between the self-reported scores and official records to establish an overall level of accuracy. In addition, the analyses targeted two additional potential group differences, tendency to deviate and actual performance levels.

To investigate the impact of direction of reported scores' deviation from the actual scores, each participant's reported value was categorized as either an overestimation, an underestimation, or as accurate. These reports were examined to identify whether the magnitude of deviation from students who overestimated and underestimated their scores differed significantly from each other.

The second grouping variable was based on actual performance levels. To examine whether low-scoring individuals inflated their scores more than high-scoring individuals in both SAT and GPA self-report values, four groups were established for each measure, using the quartile split method.

To investigate differential magnitudes of deviation based on both direction of deviation (overestimation and underestimation) and actual performance level, univariate analyses of variance were conducted on the absolute value of the deviation of the reported score from the actual score. The use of absolute values is appropriate because the direction of deviation is represented in the directional grouping variable.

Results

Students' self-reported GPA scores were found to be remarkably similar to official records. The Pearson product moment correlation revealed a significant correlation between self-reported and actual cumulative GPA, $\underline{r} = .97$, $\underline{p} < .0001$, $\underline{n} = 75$. Similarly, correlational analyses of the accuracy of the students' self-reported SAT scores revealed significant relationships between self-reports and actual performance levels for the total score ($\underline{r} = .88$, $p = .0001$, $\underline{n} = 72$), verbal subscale ($\underline{r} = .73$, $\underline{p} = .0001$, $\underline{n} = 64$), and math subscale ($\underline{r} = .89$, p

= .0001, n = 64).

To examine deviation of GPA scores, a two by four univariate analysis of variance was used, with two levels of direction of deviation (overestimation and underestimation) and four levels of actual GPA (as established by quartile placement in the sample). The ANOVA revealed a significant main effect for level of GPA on deviation from reality. Neither the main effect for direction of deviation nor the interaction produced a significant effect (see Table 1). The data indicated progressively more accurate ratings of GPA as the level of GPA increased (see Table 2 for means and standard deviations). Post-hoc analyses of group differences revealed differences between the quartiles, with the first quartile deviations being significantly higher than the third ($p < .005$) and fourth ($p < .001$), and the participants in the second quartile producing significantly higher deviations than the fourth ($p < .05$).

Table 1: Analysis of variance for group differences in deviation of self-reported GPA

Source	df	F
Direction of Deviation (D)[a]	1	.49
Performance Level (L) [b]	3	5.06***
D X L	3	1.38
Error	60	(.01)

Note: The figure in parentheses represents the mean square error. *** - $p < .001$, ** - $p < .01$, *- $p < .05$. [a] Direction of deviation includes overestimation and underestimation. [b] Performance level was determined by quartile splits on GPA.

Table 2: Deviations of self-reported scores from actual scores for GPA and SAT subscale scores

	GPA[a]			SAT Verbal			SAT Math		
Group	n	M	SD	n	M	SD	n	M	SD
Underestimation									
First Quartile	10	.117	.134	3	53.33	75.06	5	40.00	31.62
Second Quartile	8	.120	.118	4	95.00	137.23	3	13.33	5.77
Third Quartile	13	.064	.057	8	31.25	28.00	5	18.00	4.47
Fourth Quartile	11	.050	.038	4	22.50	25.00	3	40.00	20.00
Overestimation									
First Quartile	8	.221	.184	9	70.00	41.83	6	65.00	37.28
Second Quartile	9	.126	.076	4	37.50	15.00	6	36.67	20.66
Third Quartile	5	.065	.027	5	31.00	14.32	4	22.50	15.00
Fourth Quartile	4	.023	.020	3	80.00	81.85	5	36.00	19.49

Note: Reported scores are the absolute value of the difference between the actual and reported values. [a] GPA score was based on 4.0 scale.

Similar analyses were conducted on the verbal and math subscales of the SAT. Because the total score for the SAT is a combination of these two subscales, no additional analysis of the total score was conducted. To examine deviation of the SAT subscale scores, two separate two by four univariate analyses of variance were conducted, with two levels of direction and four levels of SAT performance. The ANOVA revealed no significant effects for the verbal subscale (see Table 3). The results for the math subscale revealed a trend similar to GPA, with a significant main effect for level of SAT performance (as determined by quartile placements), while the main effect for direction of deviation and interaction were not significant (see Table 3). Post-hoc analyses revealed that members in the first quartile produced significantly higher deviations than members in the second ($p < .03$) and third ($p < .004$) quartiles.

Table 3: Analyses of variance for group differences in deviation of self-reported SAT subscales

Source	df	F
SAT Verbal		
Direction of Deviation (D)[a]	1	.14
Performance Level (L) [b]	3	1.10
D X L	3	.79
Error	31	(2963.68)
SAT Math		
Direction of Deviation (D)[a]	1	2.30
Performance Level (L) [b]	3	3.66*
D X L	3	.78
Error	29	(559.48)

Note: The figures in parentheses represent the mean square error. *** - $p < .001$, ** - $p < .01$, *- $p < .05$. [a] Direction of deviation includes overestimation and underestimation[b] Performance level was determined by quartile splits on the SAT subscales.

Discussion

The results of this study of GPA and SAT self-reports allow for a general statement regarding the role of self-reported performance indicators to be made. The initial hypothesis regarding accuracy of ratings was supported, revealing that the participants had highly reliable ratings of cumulative GPA ($r = .97$). Such high correlations would suggest that overall, self-reported GPA levels are sufficiently accurate. The overall accuracies of the students' self-reported SAT scores were considerably lower than the accuracy of GPA; however, the average accuracy was still within reasonable guidelines (Nunnally and Bernstein, 1994). The results supported the expectation that the accuracy of self-reported SAT scores would be lower than self-reported GPAs.

This difference in accuracy may be related to the factors of repetition and recency. Cumulative

GPA is reported to undergraduate students on a consistent and frequent basis, typically at least two to three times per year. SAT scores, however, are not typically reported to the students once they've been admitted to the university; consequently, the majority of these participants would not likely have seen their official SAT scores for a period of two or more years.

Further investigation revealed that accuracy of self-reported scores was dependent upon the independent variable of performance level. The analyses of accuracy in self-reported GPA revealed that the bottom 25% of students provided estimates that were significantly less accurate than each of the remaining quartile groups. These data support a trend reported by Dobbins et al. (1993), who revealed that students with lower GPAs tended to inflate their scores more than students with higher averages. In a similar vein, self-reports of SAT performance generally became more accurate as actual performance increased. Overall, it appears that students at the lowest end of performance are more likely than the high-achieving groups to misrepresent their scores. This is consistent with the proposal that the students at the low-performing levels may provide inflated scores as a function of social desirability (Dobbins et al, 1993).

Contrary to the initial hypothesis, there were no differences in deviation from actual scores by those participants who overestimated and underestimated their performance levels. The expectation was that the deviations would be higher for overestimators, consistent with the social desirability hypothesis. However, no such trend was revealed, suggesting that the deviations from actual scores are due in part to errors in memory, and not all deviations are driven by a desire to misrepresent ability levels.

Given ideal conditions, there would be no sense in relying on students to report their GPA and SAT scores from memory. However, several conditions may limit a researcher's ability to gain access to official records, including administrative rules and privacy issues. When these conditions arise, forcing a researcher into a compromised methodological activity, these data suggest that researchers can rely upon self-reported GPA estimates. The data suggest that the use of self-reported SAT scores is less reliable than GPA estimations, but can be tempered by indicating to the students that accuracy is of primary interest, perhaps by assuring anonymity to the participants (see Shepperd, 1993). The use of self-reported GPA and SAT scores increases the efficiency of data collection available to researchers, particularly when these scores are simply additional variables of interest, perhaps when attempting to account for variance in designs examining course performance, test anxiety, or career orientations. The ease of acquiring these values through self-report, combined with the high levels of accuracy under the current methodology, make this practice an enticing alternative to the more laborious process of accessing official student records.

However, these results do not support the use of self-reported GPA and SAT scores for policy decisions, particularly if the students are able to determine the intent of the score collection. In situations where the students' GPA and SAT scores will be used to differentiate among candidates for selection into special programs or positions, students may be more likely to provide false estimates to improve their standing. Furthermore, this practice should not be generalized to participants at different developmental levels without assessing a pilot sample to ensure the reliability is still adequate.

References

Dobbins, G. H., Farh, J. L., and Werbel, J. D. (1993). The influence of self-monitoring and inflation of grade-point averages for research and selection purposes. *Journal of Applied Social Psychology, 23,* 321-334.

Flake, W. L., and Goldman, B. A. (1991). Comparison of grade point averages and SAT scores between reporting and nonreporting men and women and freshmen and sophomores. *Perceptual and Motor Skills, 72,* 177-178.

Frucot, V. G., and Cook, G. L. (1994). Further research on the accuracy of students' self-reported grade point averages, SAT scores, and course grades. *Perceptual and Motor Skills, 79,* 743-746.

Goldman, B. A., Flake, W. L., and Matheson, M. B. (1990). Accuracy of college students' perceptions of their SAT scores and high school and college grade point averages relative to their ability. *Perceptual and Motor Skills, 70,* 514.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory* (3rd Ed.). New York: McGraw-Hill, Inc.

Shepperd, J. A. (1993). Student derogation of the Scholastic Aptitude Test: Biases in perceptions and presentations of College Board scores. *Basic and Applied Social Psychology, 14,* 455-473.

Trice, A. D. (1990). Reliability of students' self-reports of scholastic aptitude scores: Data from juniors and seniors. *Perceptual and Motor Skills, 71,* 290.

Address all correspondence to Jerrell C. Cassady, Ph.D., Department of Educational Psychology, Ball State University, Muncie, IN 47306; jccassady@bsu.edu

Descriptors: SAT; Test Scores; Standardized tests; Methodology

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Tienken, Christopher & Michael Wilson (2001). Using state standards and tests to improve instruction. *Practical Assessment, Research & Evaluation*, 7(13). Available online: http://ericae.net/pare/getvn.asp?v=7&n=13. This paper has been viewed 2515 times since 2/23/01.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Tienken, Christopher
  Michael Wilson

Using State Standards and Tests to Improve Instruction

Christopher Tienken & Michael Wilson

Abstract:

Most states have mandated curriculum standards and tests for their students. This article describes a program used by two educators in New Jersey. The aim of the program is to help teachers understand and use their state's standards and test specifications to improve classroom instruction and raise achievement. The program is part of a research project being conducted by the authors.

Today many states around the country have curriculum standards, and state developed assessments to monitor the implementation of those standards. Most state standards define expected outcomes, that is, what students need to know and be able to do, but do not mandate specific strategies or pedagogy used by local districts. Elementary, middle and high school students around the country take at least one state mandated test during their school career. However, 35 out of 50 states do not require teachers take a course, or demonstrate competency, in the area of assessment. Hence, teachers generally have limits to their knowledge of how to design and use tests and assessment tools. Richard Stiggins (1999) wrote, " It is time to rethink the relationship between assessment and effective schooling."

It is possible for teachers and administrators to use state content and process standards, test specifications, curriculum frameworks, sample questions, educational research, and exemplar papers to improve instruction and classroom tests and assessment procedures, but limited understanding puts constraints on this use. Researchers Paul Black and Dylan Wiliam (1998) stated standards are raised only by changing what happens in the classroom, beginning with teachers and students. These researchers go on to say that a large body of evidence suggests that attention to formative assessment is a vital feature of classroom work and the development of it can raise standards.

This article describes a program used by two educators to help teachers improve instruction through a deeper understanding of state standards and test specifications. Any teacher or administrator in any state can use the process outlined in this article. Specific examples were developed using the New Jersey Core Curriculum Content Standards and that state's fourth grade mathematics test.

Developing a Knowledge Base

Understanding how standards-based state tests are constructed is the first step in being able to use them to guide and improve instruction. A test is essentially a sample of questions or activities that reflect a large body of knowledge and mental processes associated with an academic subject area. It is highly impractical to design a test that includes all of the problems that a student could ever do in each content area. Therefore, state tests are samples of possible questions from each area. All state tests are limited samples of what students are required to know in areas such as language arts, mathematics, science, etc. There are large numbers of questions that can appear on future forms of these instruments. A teacher would not be able to address all the possible questions, nor should the teacher attempt that task. However, school districts and teachers should endeavor to understand the delineation of each subject area.

School districts are under pressure to perform well on state tests and often use a test preparation strategy of giving students sample tests from commercially prepared workbooks or state released items to get ready for state tests. Although this is one strategy that can be useful for providing general information regarding student strengths and weaknesses as related to the samples, it should not be the only method used by teachers. The strategy itself, does little to educate teachers about how to use and understand state tests, standards, and test specifications. This article recommends a three-part process for developing an understanding of state assessments and using that understanding to improve instruction. That process is delineation, alignment, and calibration.

Delineation

Delineation is the first component needed to understand any standards based test. It is the process of thoroughly identifying all aspects of a particular subject domain; the aspects are also known as dimensions. Delineation involves the use of state testing documents that describe each content area of the assessment. The documents usually include test specifications, specific skill cluster information, subject area frameworks, assessment examples and exemplars, and the state standards. Delineation requires an examination of these documents for assessment dimensions such as content, cognitive level and complexity. A thorough delineation might also include analysis of the test format, motivation, the difficulty level of the questions, and related affective characteristics of the subject area.

Thoroughly examining state standards and test specifications is a way to begin delineation. The New Jersey Standards include macro or big picture statements and cumulative progress indicators that provide details about general performance expectation. The State's test specifications are particularly helpful because they go further and break the Standards down into two distinct types. Knowledge specifications describe the specific processes and content that all students must know by the end of fourth grade. Some would call these content standards. Problem solving specifications describe what students should be able to do with the content knowledge. They are also known as process standards. The following example is excerpted from the 4th grade New Jersey mathematics standards and test specification manuals.

Macro Standard 4.1: All students will develop the ability to pose and solve mathematical problems in mathematics, other disciplines, and everyday experiences.

> Cumulative Progress Indicator 4.1.2: Recognize, formulate, and solve problems arising from mathematical situations and everyday experiences.
>
> > Test Specification Manual - Cluster IV Discrete Mathematics:
> >
> > Knowledge (content standards): Students should have a conceptual understanding of: Tree diagram
> >
> > Problem Solving (process standards): In problem solving settings, students should be able to: Draw and interpret networks and tree diagrams

After reviewing the 4th Grade New Jersey Core Curriculum Content Standards and test specifications for mathematics, a teacher would be able to identify seven distinct mathematics strands or dimensions. Those strands are Numeration and Number Theory, Whole Number Operations, Fractions and Decimals, Measurement/ Time/ Money, Geometry, Probability/Statistics, and Pre-algebra. Figure 1 represents the content delineation of the domain of mathematics after a team of 4th grade teachers examined the New Jersey Core Curriculum Content Standards, 4th grade state test specifications, and the local curriculum.

Mathematics Domain

Numeration/Number Theory

Fractions/Decimals

Geometry

Probability/Statistics

Whole Number Operations

Measurement/Time/Money

Pre-algebra

(Delineated Strands / Dimensions)

(Figure 1 –A delineation of the domain of Mathematics)

Working through the different dimensions associated with the delineation process helps to increase teacher and administrator understanding of each content area and its relationship to the standards, classroom instruction and assessment.

The following activities can begin once teachers and administrators specify all of the subject area dimensions:

- selecting and designing classroom assessments and practice questions
- revising and designing curriculum that is congruent with the content identified in the state standards and the district's delineation of the state designed exams
- designing teacher training using instructional techniques that support these dimensions

A closer look at the 4th grade New Jersey Core Curriculum Content Standards and test specifications for mathematics reveals an emphasis on performance and the use of mathematics to solve open ended and word problems. The test specifications for that exam imply that the mathematics test questions are primarily composed of problem solving tasks. Therefore, it is safe to assume that test questions will require thinking in the application, analysis, and perhaps synthesis and evaluation levels of cognition.

Alignment

During the alignment phase, administrators and teachers work to identify, analyze, generalize, and describe the links between the various elements associated with the subject area previously delineated and the sample questions selected for practice or classroom activities to assess student progress. The sample questions and student assessments can be derived from several sources including state released test items, commercially manufactured test preparation materials, or teacher made activities. Teachers and administrators examine linkages in the materials, organization, textbooks, instructional strategies and other elements described in the curricula and used in daily instructional activities to ensure consistency with the district's delineation of the state assessment.

Using and understanding the test specifications become even more important at this stage. Let's imagine that a pair of 4th grade teachers recently completed a delineation of the mathematics domain and identified their next unit of study. The unit centered on Standard 4.1.1 and the test specification listed below. Reviewing the prior example from the test specification manual and Cluster IV the teacher would complete several alignment tasks:

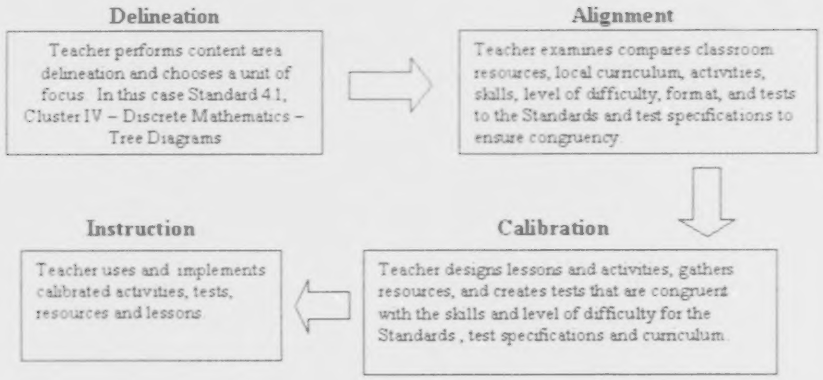
Test Specification Manual - Cluster IV Discrete Mathematics:

> Knowledge (content standards): Students should have a conceptual understanding of: Tree diagram
>
> Problem Solving (process standards): In problem solving settings, students should be able to: Draw and interpret networks and tree diagrams
>
> Tasks:
>
> 1. Review classroom resources, curriculum, textbooks, teacher activities, student thinking strategies and tests to ensure that the above test specifications and macro standard are addressed on the knowledge and problem solving level. Do the teacher resource materials and classroom instruction address the proper skills?
> 2. Review the above factors to ensure congruency between the level of difficulty required by the standards and specifications, and the difficulty of the actual teacher resources and activities. Do the teacher's tests, lessons, activities etc., match the difficulty level required by the standards and specifications?
> 3. The teacher must also consider format. Although less important than skills and difficulty, the teacher resources, activities, and tests should familiarize the students with state test question formats.

Teachers must align classroom assignments and activities to the subject area delineation to ensure congruency.

Calibration

After completing the delineation and beginning the alignment processes, calibration begins. Calibration is the act of conducting communications and interactions with teaching staff based on the information identified in delineation and used in alignment. The calibration process ensures that the conceptualization of content, cognitive process, complexity, formats, etc. is consistently understood for each subject area. Calibration, in its simplest form, is designing classroom instruction, activities and assessments that are congruent with content area delineation and alignment. Using the prior mathematics vignette as an example, one can begin to see how the process takes place. Figure 2 represents the sequence of events leading up to calibration.



Teacher performs content area delineation and chooses a unit of focus. In this case Standard 4.1, Cluster IV – Discrete Mathematics – Tree Diagrams

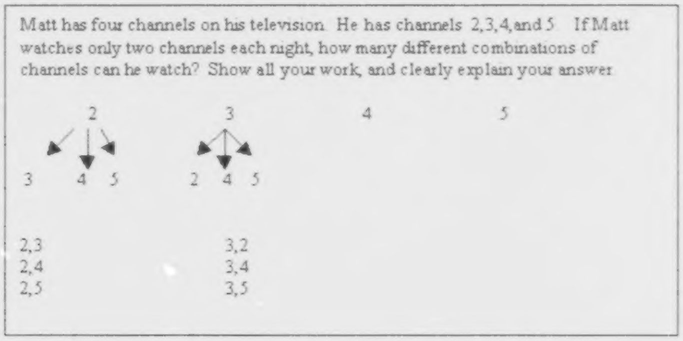
Teacher examines compares classroom resources, local curriculum, activities, skills, level of difficulty, format, and tests to the Standards and test specifications to ensure congruency.

Teacher designs lessons and activities, gathers resources, and creates tests that are congruent with the skills and level of difficulty for the Standards , test specifications and curriculum.

Teacher uses and implements calibrated activities, tests, resources and lessons.

(Figure 2. Delineation, Alignment, and Calibration Flow of Events)

Imagine that a 4th grade teacher completed delineation and alignment and discovered that her/his program was missing a unit on discrete mathematics. That teacher would develop objectives related to understanding, using, and interpreting tree diagrams. Figure 3 is a sample activity / test question created by 4th grade teacher Terry Maher to begin addressing the aspect of discrete math noted in the Cluster IV test specification.



Matt has four channels on his television. He has channels 2,3,4,and 5. If Matt watches only two channels each night, how many different combinations of channels can he watch? Show all your work, and clearly explain your answer.

(Figure 3: A sample activity / test question)

Calibration is any action that helps teachers design activities and construct assessments based on the dimensions of state assessments and standards. This process helps to foster a collective understanding and agreement of the dimensions and domains of each content area. It should be a team effort based on group inquiry.

Using Score Reports to Improve Calibration

As teachers gain a better understanding of how student work reflects the standards and test specifications through delineation, alignment and calibration, their efficiency and accuracy at identifying which students are meeting the standards should increase. Herein lies the usefulness of score reports. State test score reports sort students into categories of varying proficiency. For example, a student who scores partially proficient, proficient, or advanced proficient on a state language arts test may also show some congruency in the level of achievement in his/her well-aligned school work and classroom assessments. As teachers become better calibrated, they will be able to answer questions such as: Is the student showing partial proficiency, proficiency, or advanced proficiency on class assessments? If not, why? Is the difficulty level of the class work comparable to the state exam? What can I do to help this student meet the state standards? Is my program meeting the standards?

Predicting Outcomes

Teachers can reflect upon their level of calibration accuracy by attempting to predict student results on state assessments. This type of exercise acts as an extension to the calibration process and can provide teachers with a way to get a very general sense of their level of calibration. Teachers should be aware that there would not be 100% agreement between a student's performance on well-calibrated classroom tests and state assessments based on many factors of test design. This process is meant to compliment the calibration exercises and provide the teacher with extra data.

To begin the prediction process, the teacher uses a list of the students taking the test. Beside each name, the teacher enters a predicted score level. When the state assessment scores arrive, the teacher can compute the level of accuracy as shown below.

Name	Prediction	Score
Allan	Proficient	Adv. Proficient
Ann	Proficient	Proficient
Tamika	Adv. Proficient	Proficient
Bronson	Partial Proficient	Partial Proficient

The list above shows a 50% level of success in the predictions made. The teacher making the predictions returns to each student's work and compares the successful predictions with the unsuccessful ones to gain a better idea of how the assessment performances reflect the aligned student work. Student work associated with actual test scores can form the basis for subsequent calibration discussions. Student work connected to state assessment score levels can also function as scoring examples that students refer to when judging their own work.

Final Thoughts

The process outlined in this paper is very different from the idea of using testing materials and example tests to teach specific items on state assessments. Although there is a place for such strategies, this article suggests that it is more important for the teacher to understand the entirety of each subject area, and where state test content fits within each of these areas. Teachers must teach toward an understanding of the subject areas while they align and calibrate their classroom activities, resources, tests, and instruction with the specifications and skills required by each state's standards. There is a distinct difference between traditional notions of test preparation and aligning and calibrating instruction and assessments with the content, cognition, difficulty, and format of state assessment instruments, specifications, and standards. The aim is to ensure that teachers understand, and calibrate their classrooms with respect to the entire process and do not simply focus on how to answer specific types of test questions.

The questions will change, but the underlying skills and concepts will not. One must be careful not to wallow in the mire of test prep. As educators, we are trying to link the classroom activities to the standards and skills set by the state. Delineation, alignment, and calibration are academic endeavors that demand unending commitment. Do not expect to accomplish alignment or calibration at an in-service day, or even during the course of a school year. This ongoing process requires constant attention. The administration must provide the time and resources to conduct frequent calibration meetings to examine such things as classroom work and student assessment samples. Beware, it is easy to fall out of alignment and calibration and into test prep.

References

Black, Paul, and Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment, *Phi Delta Kappan* October, pp. 139-148.

Stiggins, R. (1999). Assessment, student confidence, and school success, *Phi Delta Kappan* November, pp. 191-198.

Related Works and Readings

Neill, D. (1997 September). Transforming student assessment, *Phi Delta Kappan,* pp.35-36.

Sadler, D. (1989). Formative assessment and the design of instructional systems, *Instructional Science,* vol. 18, 1989, pp. 119-44.

Schafer, W. & Lissitz, R. (1987). Measurement training for school personnel: Recommendations and reality, *Journal of Teacher Education,* vol. 38. 3, pp. 57-63.

Stiggins, R. & Conklin, N. (1992). In teachers' hands: Investigating the practice of classroom assessment. Albany: State University of New York Press.

Christopher Tienken is the Curriculum Coordinator for the Absecon School District in New Jersey. His responsibilities include curriculum, instruction, professional development, and developing the district's assessment system. He is a member of the Epsilon Zeta Chapter of Kappa Delta Pi. He can be reached at goteach1@hotmail.com

Dr. Michael Wilson is the Director of Assessment and Curriculum for the Plainfield School District in New Jersey. He actively conducts research in the field of testing and assessment and was the manager of several New Jersey statewide assessments. He can be reached at drmikewilson@yahoo.com

Descriptors: Performance Based Assessments; State Assessment; Alignment; Academic Standards

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Rudner, Lawrence M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7(14). Available online: http://ericae.net/pare/getvn.asp?v=7&n=14. This paper has been viewed 1002 times since 3/30/01.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

Find articles in ERIC written by
Rudner, Lawrence M.

Computing the Expected Proportions of Misclassified Examinees

Lawrence M. Rudner
LMP Associates & the Maryland Assessment Research Center for Education Success

Unless a test is perfectly reliable and thus contains no error, every time we make a classification based on a test score, we should expect some number of misclassifications. We can expect that some examinees whose true ability is above the cut score will be incorrectly classified as non-masters (false negatives) and that some number of low-ability examinees will be incorrectly classified as masters (false positives).

This paper provides and illustrates a method to compute the expected number of misclassifications. This information can help policy makers decide whether the risks are sufficiently small or whether the costs for improvements are justified. One particularly useful application of this procedure is to estimate the probability of a true master consistently failing multiple test administrations. Another is to examine the impact of cut score adjustments. Web-based software to apply this method is available at http://ericae.net/pare/misclass/class.asp. PC based software is available from the author.

Approach

I will develop the procedure using three-parameter item response theory and two state classifications (mastery and non-mastery). There are classical test theory analogs and the logic can easily be extended to more categories. I start with a test that maps individual scores (θ_1's) onto a continuous scale (a θ scale) and a cut score on that scale (θ_c) used to classify examinees into one of two discrete categories. Examinees whose scores are above the cut score will be classified as masters; those below as non-masters.

We make a distinction between the categories examinees should be placed in based on their true scores and the categories they are placed in based on observed score. The goal of this paper is to create, and then analyze, a two by two classification table, such as Table 1, indicating the expected proportions of correct and incorrect classifications.

Table 1: Classification table	
classified master true master	classified non-master true master
classified master true non-master	classified non-master true non-master

In table 1, the upper left and lower right quadrants represent correct classifications; the other two quadrants represent incorrect classifications.

The expected proportion of all examinee classified as a masters that are true non-masters is:

$$\mathrm{P(cm,n)} = \sum_{\theta_i < \theta_c} P(\hat{\theta} > \theta_c | \theta_i) f(\theta_i) / n \qquad (1)$$

It should be noted that the phrases *true masters* and *true non-masters* are statistical terms meaning that the true ability is above or below the arbitrarily set cut score. In (1), $P(\hat{\theta} > \theta_c \mid \theta_1)$ is the probability of having an observed score, $\hat{\theta}$, above the cut score given a true score equal to θ_1, $f(\theta_1)$ is the expected number of people whose true score is θ_1 and n is the total number of examinees. Thus $P(\hat{\theta} > \theta_c \mid \theta_1) f(\theta_1)$ is the expected number of people whose true score is θ_1 that will be classified as masters, i.e., will have an observed score greater than the cut score. We sum this value over all examinees whose true score is less than the cut score and divide by n to obtain the probability of being misclassified as a master (false positive).

Similarly, the expected proportion of false negative is:

$$\mathrm{P(cn,m)} = \sum_{\theta_i > \theta_c} P(\hat{\theta} < \theta_c | \theta_i) f(\theta_i) / n \qquad (2)$$

Referring to (1), the probability of having an observed score above the cut score given θ_1, $P(\hat{\theta} > \theta_c \mid \theta_1)$, is the area under the normal curve and to the right of

$$z = \frac{\theta_c - \theta_i}{se(\theta_i)} \tag{3}$$

This is illustrated in Figure 1. The taller bell curve represents the distribution of ability in the entire population and the cut score is set at $\theta_c = -.2$. The smaller curve represents the expected distribution of observed values of theta for examinees with a true value of $\theta_1 = -.5$. Examines with a true score of $\theta_1 = -.5$ are non-masters and should be classified that way. However, the observed scores will vary around $\theta_1 = -.5$. The shaded area to the right of the cut score represents the proportion of examinees whose true score of -.5 that can be expected to be misclassified as masters. Figure 1 is for just one value of theta. To determine P(cm,n), one would have curves for each value of theta less than θ_c.

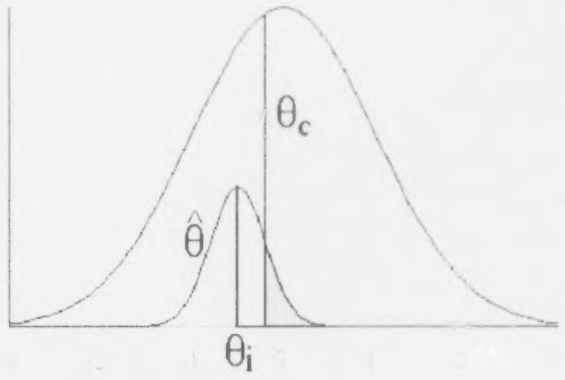


Figure 1: False positives for examinees at one ability level.

In (3), $se(\theta_1)$, the standard error of measurement evaluated at a score of θ_1, is

$$se(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}} \tag{4}$$

where $I(\theta_1)$ is the test information function evaluated at a score of θ_1. Lord(1980, pages 72-74) provides equations for $I(\theta_1)$ using weighted composite scoring, number right scoring, and item response theory scoring. Using IRT, the test information function at θ_1 is the sum of the item information functions at θ_1 which can be evaluated from the IRT *a*, *b*, and *c* item parameters.

The expected proportions of examinees whose true score is θ_1, $f(\theta_1)/n$, can be estimated from the pilot sample. Denoting the probability of obtaining a score of θ_1 as $P(\theta_1)$, and looking at θ as a continuous rather than a discrete variable, (1) and (2) become:

$$P(\mathrm{cm,n}) = \int_{\theta_1=-\infty}^{\theta_c} P(\hat{\theta} > \theta_c|\theta_i)P(\theta_i)\,d\theta_i \qquad (5)$$

$$P(\mathrm{cn,m}) = \int_{\theta_1=\theta_c}^{\infty} P(\hat{\theta} < \theta_c|\theta_i)P(\theta_i)\,d\theta_i \qquad (6)$$

To complete the set of equations needed for our two-by-two classification table, the probability of correctly being classified as a master and the probability of correctly being classified as a non master are

$$P(\mathrm{cm,m}) = \int_{\theta_1=\theta_c}^{\infty} P(\hat{\theta} > \theta_c|\theta_i)P(\theta_i)\,d\theta_i \qquad (7)$$

$$P(\mathrm{cn,n}) = \int_{\theta_1=-\infty}^{\theta_c} P(\hat{\theta} < \theta_c|\theta_i)P(\theta_i)\,d\theta_i \qquad (8)$$

If one assumes a normal distribution of mean μ and standard deviation , then $P(\theta_1)$ is the height of the Gaussian curve evaluated at $(\theta_1 - \mu)/\sigma$.

Illustration

To illustrate an application of the above formulas, I generated a data set consisting of the item parameters for 50 items. The *a, b,* and *c* parameters were each normally distributed with means of 1.43, 0.00, and 0.20 respectively. Assuming a normally distributed examinee population with mean=0 and sd=1.0, the weighted average standard error is .393 which corresponds to a classical reliability coefficient of .92.

Applying equations (5) through (8) with a passing score of $\theta_c = -.25$, which is about the 40th percentile, yields

Table 2: Expected and true examinee classifications

		Expected Classification	
		Master	Non-Master
True	Master	53.9	5.0
	Non-Master	6.5	34.6
		Accuracy = 88.5%	

This is a highly reliable test that is expected to accurately classify examinees an impressive 88.5% of the time. On closer examination however, 5.0/(5.0+53.9)= 8.5% of the true masters can be expected to be incorrectly classified as non-masters and 6.5/(6.5+34.6)=15.8% of the non-masters can be expected to be incorrectly classified as masters. Whether this is sufficiently accurate is matter of judgment.

The percent of expected misclassifications can be used as the benefit side of a cost benefit analysis. We can examine the potential gain obtained by increasing the quality of the test by simulating the addition of items with peak information at the cut score. If we add 10 items with parameters a=2.0, b=-.25 and c=0, we obtain the following classifications:

Table 3: Expected and true examinee classifications adding 10 optimal items

		Expected Classification	
		Master	Non-Master
True	Master	55.4	3.6
	Non-Master	4.8	36.3
		Accuracy = 91.7%	

Accuracy goes up from 89.1% to 91.7% and the proportion of false negatives goes down from 5.0 to 3.6% of all test takers. Here, the marginal benefits of improving the test may not justify the associated costs. On the other hand, this is a 28% reduction in the number of false positives (5.0-3.6)/5.0. This could be very worthwhile if false negatives are costly.

A common approach to minimizing the number of false negatives is to render due process by providing repeated opportunities to demonstrate mastery. If we make the convenient assumption that test scores from different administrations are independent, then the probability of a true master being misclassified after three attempts is the product of the probabilities or, in this case, $.085^3$=.0006. With three opportunities to pass, only a very small fraction of true masters will be misclassified as non-masters. Of course, the probability that a non-master will pass in three tries increase from .158 to .158+.158*(1-.158)+.158*(1-.158(1-.158))= .428.

One common, albeit often misguided, approach to reducing the number of false negatives is to set the operational cut score to a value lower than that recommended by a standard setting panel. This can be modeled by maintaining the original cut score in equations (5) through (8), but integrating to and from the adjusted cut score rather than the original cut score. Using an operational cut score of -.40 rather than -.25 yields the following results:

Table 4: Expected and true examinee classifications using an adjusted operational cut score

		Expected Classification	
		Master ($\theta_{c'}>-.40$)	Non-Master ($\theta_{c'}<-.40$)
True	Master ($\theta_c>-.25$)	56.2	2.7
	Non-Master ($\theta_c< -.25$)	9.8	31.3
Accuracy = 87.5%			

Now the true master is half as likely to be misclassified, 2.7/(2.7+56.2)= 4.6%. However, the non-master now has a 9.8/(9.8+31.3)=23.8% chance of being classified as a master. If the original standard were meaningful then setting a lower operational cut score is a poor alternative. If the rationale for lowering the operational test score is to recognize the error associated with assessment, then the approach is misguided. Error is assumed to be normally distributed. An individual's score is as likely to be above his or her true score as it is to be below. It would, however, be appropriate to make an adjustment in order to recognize the error associated with the standard setting practice. This could be viewed as simply implementing the judgment of a higher authority.

Concluding remarks

This paper consistently talked about true masters and true non-masters. One must recognize that the classifications always involve judgment (Dwyer, 1996) and that, despite the use of quantitative techniques, cut scores are always arbitrary (Glass, 1978). We cannot say that a person has mastered Algebra just because his or her true or observed score is above some cut point. Algebra, or almost any domain, represents a collection of skills and hence is not truly unidimensional. Because we are talking about a multidimensional set, it is illogical to talk about mastery as if it were a unidimensional set. The only true masters are those who get everything right on the content sampled from the larger domain.

Nevertheless, we recognize the need to establish cut scores; mastery in this paper refers to people who score above some established cut score. When that mastery -

nonmastery decision affects real people, then the expected impact of that decision should be examined. This paper provides a way to estimate the number of false positives and false negatives using 1) the standard error, which could be the standard error of measurement or an IRT standard error, and 2) the expected examinee ability distribution, which could be estimated from a pilot sample or based on a distribution assumption, such as a normality assumption. It is our hope that this tool will lead to better, more informed, decision making.

Notes:

Internet-based software to apply this technique is available at http://ericae.net/pare/misclass/class.asp. Comparable QuickBasic source code and a Windows executable file are available from the author.

This research was sponsored with funds from the National Institute for Student Achievement, Curriculum and Assessment, U.S. Department of Education, grant award R305T010130. The views and opinions expressed in this paper are those of the author and do not necessarily reflect those of the funding agency.

References and additional reading

Cizek, Gregory J. (1996). An NCME Instructional module on setting passing scores. *Educational Measurement: Issues and Practice*, 15(2), 20-31.

Dwyer, Carol Anne (1996). Cut scores and testing: statistics, judgment, truth, and error. *Psychological Assessment*, 8(4), 360-62.

Geisinger, Kurt F. (1991).Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice*, 10(2), 17-22.

Glass, Gene V (1978).Standards and criteria.*Journal of Educational Measurement*, 15(4), 237-61.

Lord, Frederick M. (1980). *Applications of item response theory*. New Jersey: Lawrence Erlbaum and Associates.

Lawrence M. Rudner is the Director of the ERIC Clearinghouse on Assessment and Evaluation and the Assistant Director of the Maryland Assessment Research Center for Education Success, Department of Measurement Statistics and Evaluation, 1129 Shriver Laboratory, University of Maryland College Park, MD 20742.

Descriptors: Classification; *Cutting Scores; *Error of Measurement; Pass Fail Grading; * Scoring; * Statistical Analysis

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Schafer, William D. (2001). Replication in field research. *Practical Assessment, Research & Evaluation*, 7(15). Available online: http://ericae.net/pare/getvn.asp?v=7&n=15. This paper has been viewed 998 times since 4/18/01.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

Find articles in ERIC written by
Schafer, William D.

Replication in Field Research

William D. Schafer
University of Maryland

This article suggests the routine use of replications in field studies. Since replications are generally independent, it is usually possible to synthesize them quantitatively using meta-analysis, a technique heretofore associated primarily with amalgamating prior work. It is argued that the use of replication as a feature in data collection and quantitative synthesis for data analysis is especially attractive for those investigators whose research paradigm choices are limited because they are working in field environments. Two examples are described briefly.

Control of extraneous variables is a fundamental condition to causal interpretations of research (Johnson, 2001). Randomization of participants to treatment conditions has long been considered a powerful method of control, so much so that this is the distinguishing characteristic between true experimental and other types of research (Campbell & Stanley, 1963). When a researcher uses randomization, it is clear that the basis upon which participants receive treatment conditions is unrelated except by chance to any variable that can be confounded with the treatments.

A great deal of research is done in field settings in education. State-level or district-based researchers, for example, are often interested in practical interventions that can occur naturally in schools. However, randomization is typically unavailable to those who work in field settings because the investigator is not able to manipulate treatment conditions at the level of the individual participant. This often arises because institutions such as schools are reluctant to move participants (e.g., students) from group to group (e.g., class to class) or otherwise assign them to groups according to researcher needs. Similarly, it may not be possible even to determine randomly which group receives which treatment condition, that being decided through other means, such as teacher choice.

Failing randomization, one approach used in the field is to measure extraneous variables and employ statistical control (e.g., analysis of covariance). Pedhazur (1997) describes three common contexts for statistical control with intact groups: attempting to equate them on the outcome variable(s) using one or more pretest(s), attempting to control for other variable(s) in looking at mean differences, and attempting to control for other variable(s) in looking at differences in regressions. He points out that these are usually invalid uses of analysis of covariance.

Because statistical procedures are generally less effective than experimental control, theoretical inferences about relationships observed in field settings are often subject to multiple reasonable internal validity threats. And in many cases it is not even possible to measure extraneous variables effectively, such as when limited time is available, when the number of participants in the research is limited, or when the measurement is too intrusive. Johnson (2001) has recently concluded that there is little that can be gained from a single, non-experimental research study. A feasible alternative that can enhance the ability of field investigators to draw causal inferences in field settings clearly would be an advantage.

In field contexts, there are typically many opportunities available to investigators that are not open to researchers in more controlled settings. Laboratory researchers commonly have small pools of potential participants to select from and may need to expend nontrivial resources to obtain their cooperation. On the other hand, in applied settings such as classrooms and schools, and especially for employees of the institution, students or other participants are often generously available as long as the intrusion of the research is minimal. Many investigators in the field thus have broad feasible research opportunities that laboratory researchers do not enjoy. It is therefore possible in common applied research settings to be able to repeat, or replicate, a study design more than once.

It is argued here that careful planning of replications can enhance the interpretability of applied research. When results are consistent across several studies, there is a stronger basis for observed relationship(s) than the support that is available within each study by itself, since results that have been replicated are considered more likely to generalize (continue to be observed). It is also possible to compare the studies with each other to identify constructs that interact with, or moderate, relationships. Although these advantages exist whether or not the research includes experimental control, the opportunity to replicate a basic study design in multiple field contexts is more likely to be available to the applied researcher and is a technique that can lead to stronger inferences in any setting. Thus, it is recommended that persons who conduct field research try to include replication as a fundamental feature in their studies.

The analysis of the several studies' results should also be addressed. Meta-analysis is an attractive vehicle for combining, or synthesizing, a series of research replications. Although meta-analysis is generally thought of as a means for studying an existing research literature quantitatively, it also may be used to analyze a series of related studies generated within a single project.

In the remainder of this article, pertinent features of meta-analysis are discussed briefly and then two examples are described in which multiple replications of a basic field design have been analyzed using meta-analysis to strengthen the evidence available. The basic designs differ markedly in the two examples. Finally, some design approaches for applied researchers thinking about using replications are discussed.

Meta-Analysis

Meta-analysis is commonly used to synthesize the findings of multiple, but related, research studies. Those who are unfamiliar with meta-analysis can find a brief overview along with a completely analyzed example in Schafer (1999). More extensive discussions on a broad array of topics pertinent to meta-analysis are widely available in Hedges & Olkin (1985) and Cooper & Hedges (1994).

Fundamental to meta-analysis is an effect-size measure calculated within a study. An effect-size measure may be used to compare two groups or to relate two variables. For example, the difference between two group means divided by the pooled standard deviation of the two groups in a study might be the effect-size measure [when adjusted for bias, this is Hedges & Olkin's (1985) d index]. Another might be the correlation between two variables in a study. In general, to be used in a meta-analysis, an effect-size measure must be capable of transformation to a normally distributed statistic with a known variance. Under reasonable assumptions, both the examples here are appropriate.

Techniques are described in the cited sources that allow a researcher to model the size of the effect (the effect-size index) as a result of study characteristics. That is, equations may be written, as in multiple regression, for relationships between study characteristics as predictors and an effect-size index as the criterion. These study characteristics may be descriptive of the participants, of the settings, of the treatment implementations, or of the outcome variables; in other words, virtually anything that can differentiate studies from each other can be used in the analysis as study characteristics.

In one typical approach to meta-analysis, an effect-size index is calculated for each study. The suitably weighted average of the effect sizes is tested against a null hypothesis of zero. Variation of the studies' effect sizes about the average is tested to determine whether it is at a greater-than-chance level and, if it is, then a study characteristic may be entered into a model (equation), so that effect size is then predicted as the sum of a constant (intercept) and a study characteristic scaled with (multiplied by) a slope estimate. The slope estimate is tested against the null hypothesis of zero. The variance of the effect sizes about the model (the residuals) is compared with the chance level. If homogeneity (chance-level variance) is achieved, modeling ceases; otherwise further study characteristics are added to the model. Of course, variations exist, some as solutions to special problems that may arise; only a very (over)simplified treatment is described here.

Example 1: Descriptive Gains for Schools

A descriptive, or non-experimental, design is one in which there is no manipulation of treatments. The research problem studied in Guthrie, Schafer, Von Secker, & Alban (2000) was the relationships between instructional characteristics of schools and the variation they showed in their degrees of gain or loss in student achievement over a year's time (growth). The effect-size index was the bias-corrected difference between school means at a target grade level between year one and year two on a statewide, standardized test, divided by the pooled standard deviation for the two years. The indexes were scaled so a positive difference showed improvement. The study was replicated in all six tested content areas at both tested grade levels in all 33 schools in three volunteer districts for a total of 396 effect sizes.

The independent variables in the meta-analysis were school means for teacher-reports of emphasis devoted to different approaches in reading instruction. All teachers in each school were surveyed on a questionnaire with six subscales that had been developed through factor analysis using data from a fourth volunteer district in an earlier study.

The meta-analyses were used to evaluate the association of the set of six instructional variables to achievement growth, of each variable individually to growth, and of each variable as a unique predictor of growth in a six-predictor model. The six content areas at each of the grade levels were analyzed separately. The results of the syntheses were interpretable and generally consistent with an extensive literature review for these variables.

Although it is statistically possible to compare the two years of data for any one school, that single finding by itself would not have been remarkable. While the school might have developed instructional hypotheses for the direction and degree of growth observed, there would have been far too many plausible competing explanations for the difference, such as teacher turnover, test form calibrations, and student aptitude, for example. While the replicated study cannot entirely substitute for experimental control through randomization, the plausibility of at least some of the rival explanations is decreased if instructional explanations can be observed across replications, as they were in this example study. Indeed, only by replicating the fundamental growth-study design was it possible to study the instructional characteristics of the schools as variables used to explain differences among gains across schools.

Example 2: Static Group Comparisons

A static group comparison design is one in which intact groups are randomly assigned to treatments (Campbell & Stanley, 1963). Schafer, Swanson, Bené, & Newberry (2001) studied the effects on student achievement of a treatment consisting of a workshop for high school teachers centering on an instructional method (use of rubrics). Districts nominated teacher pairs within content areas, with the classes for the two members of a given pair chosen to consist of students with similar abilities. There were 46 teacher pairs who provided complete data, evenly divided among four instructional content areas (92 teachers and 3,191 students supplied useable data in the study).

The two teachers in each pair were randomly assigned by coin flip to treatment or control conditions. The treatment, attended by one teacher from each pair, was the experimental manipulation. At the end of the study's duration, each student received a test consisting of two parts, a selected-response section and a constructed-response section. The nature of the study suggested that these two parts might yield different results and so effect sizes were calculated separately for each of the item formats. Each effect size was the difference between the means of the two classes divided by the pooled standard deviation and scaled such that a positive effect size favored the treatment.

This study was part of a larger study that required more than one form of the test. Accordingly, there were three forms in each content area. They were distributed randomly in each classroom, yielding six effect sizes (two formats on each of three forms) for each of the 46 teacher pairs, or 276 effect sizes across the four content areas.

Although there were six non-equated test forms in each of four distinct content areas, it was possible to synthesize the results of these disparate conditions in one analysis and to differentiate the findings in a planned way by contents and by forms. An interpretable pattern of outcomes was obtained and related to prior literature.

In general, there are too many competing plausible rival explanations for observed achievement differences between the two intact groups for this study's single-replicate design, in isolation, to be interesting as evidence for a difference between the instructional methods. But by using replications it was possible to synthesize findings from multiple parallel studies and thus to enhance the ability to draw inferences from the overall results.

Discussion

Consistent with Johnson's (2001) suggestions for strengthening interpretations of causality from non-experimental research, this article has recommended planning replications in field settings. The examples illustrate ways in which these replicated field designs can be synthesized to enhance the inferences that can be drawn from them. Further, when planned replications are used, it is possible also to plan for the measurement of variables that should prove useful to model effect sizes in a meta-analysis (e.g., the instructional variables in example 1). Fortunately for the researcher, a meta-analysis based on planned replications is far more straightforward to implement than a traditional synthesis of a disparate literature since fewer challenges, such as design differences, inadequate information, and inconsistent reporting of results across studies, exist.

An investigator planning to use replications in field research must make several decisions. Some of these are discussed below.

The basic design. The stronger the basic design, the stronger the inferences that may be made from any one replicate, and thus from the overall meta-analysis. The strongest feasible design should be chosen. Cook and Campbell (1979) provide an overview of designs that are particularly suitable in applied research contexts and discuss their strengths and weaknesses. It is important to be very clear what variable is independent and what is dependent in the basic design. In the two examples here, the independent variable was time (year 1 vs. year 2) in the first and presence or absence of the instruction workshop in the second. In both, the dependent variable was achievement. While year could not be manipulated in the first (the basic design was non-experimental), it was possible to manipulate the workshop in the second. Random assignment of instructors to workshop conditions strengthened that study [the basic design was pre-experimental (Campbell & Stanley, 1963)].

The effect-size measure. Magnitude of effect should be capable of coding as a standardized measure indicating direction and strength of relationship between the independent and dependent variables. Its quantification should yield an index that is is normally distributed and has a known or estimatable variance. Rosenthal (1994) provides a menu of possibilities. Three common examples that differ depending on the scaling of the two variables are: both continuous (the correlation coefficient, r); both dichotomous (the log-odds ratio, L); or, as in the two examples here, the independent variable a dichotomy but the dependent variable continuous (bias-corrected d, discussed above).

Maintaining effect-size independence. The effect sizes are assumed to be independent in a meta-analysis. That is generally the case across studies, but is not always true within studies. In our two examples, each study produced several dependent effect-size indices. Dependencies created by the measurement of six content areas in each school were ignored in the first study by analyzing each grade level and content area separately; in the second study, the six tests were analyzed together at first and a Bonferroni-like correction was applied throughout the analyses (Gleser & Olkin, 1994). Of course, care should be taken in field studies that the sites at which the replications occur maintain separation; sharing of information by participants across replications can threaten effect-size independence.

The variables to be measured. Besides the independent and dependent variables, it is advantageous to capitalize on the opportunity to measure variables that could be related to effect size (study characteristics). To generate a list of these, the researcher might consider how he or she might explain any observed differences that could appear among effect sizes across replicates. Whether substantive or artifactual, those explanations virtually always will be based on variables that should, if possible, be measured. These could be different contexts and dependent variables as in our second example in which effect sizes yielded by four different content areas and two test formats were combined into one meta-analysis. Or they may be descriptive of persons, such as demographics or aptitudes, or settings such as physical features in schools or classrooms. Coding characteristics of the replications that produced the different effect sizes provides data that are analyzed through relating these characteristics as independent variables to the effect sizes as dependent variables in the meta-analysis. The potential for assessing study differences that may be related to magnitude of effect represents an opportunity for creativity in designing robust multiple-study investigations through replication.

Meta-analysis is a relatively new approach to data analysis and the field is changing rapidly. One recent advance has been development of effective methods to conduct random-effects model analyses. Hedges & Vivea (1998) present a straightforward and relatively simple modification that is consistent with the techniques used in the two examples cited here. They also provide a worked example. An advantage of using a random model is that the results generalize to a population of studies not included in the present analysis, whereas in the two examples described here, the conclusions were restricted to the specific replications themselves. Hedges & Vivea (1998) discuss the conditions under which each type of analysis, fixed or random, is more appropriate.

References

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

Cooper, H. & Hedges, L. V. (1994). *The handbook of research synthesis.* New York: Sage.

Gleser, L. J. & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339-355). New York: Sage.

Guthrie, J. T., Schafer, W. D., Von Secker, C., & Alban, T. (2000). Contributions of instructional practices to reading achievement in a statewide improvement program. *Journal of Educational Research, 93,* 211-225.

Hedges, L. V. & Olkin, I (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hedges, L. V. & Vivea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486-504.

Johnson, B. (2001). Toward a new classification of nonexperimental quantitative research. *Educational Researcher, 30*(2), 3-13.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd Ed.). Orlando, FL: Harcourt Brace.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Sage.

Schafer, W. D. (1999). An overview of meta-analysis. *Measurement and Evaluation in Counseling and Development, 32,* 43-61.

Schafer, W. D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education, 14,* 151-170.

William D. Schafer is an Affiliated Professor with Emeritus status, in the Maryland Assessment Research Center for Education Success, H. R. W. Benjamin Building Room 1230, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD 20742-1115.

Descriptors: Meta-analysis; replication; test scores; field research; research methodology

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▼ ☐ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Ding, Cody S. (2001). Profile analysis: multidimensional scaling approach. *Practical Assessment, Research & Evaluation*, 7(16). Available online: http://ericae.net/pare/getvn.asp?v=7&n=16. This paper has been viewed 1079 times since 4/27/01.

- Find similar papers in
  - ERICAE Full Text Library
  - *Pract Assess, Res & Eval*
  - ERIC RIE & CIJE 1990-
  - ERIC On-Demand Docs
  - ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  - Ding, Cody S.

Profile Analysis: Multidimensional Scaling Approach

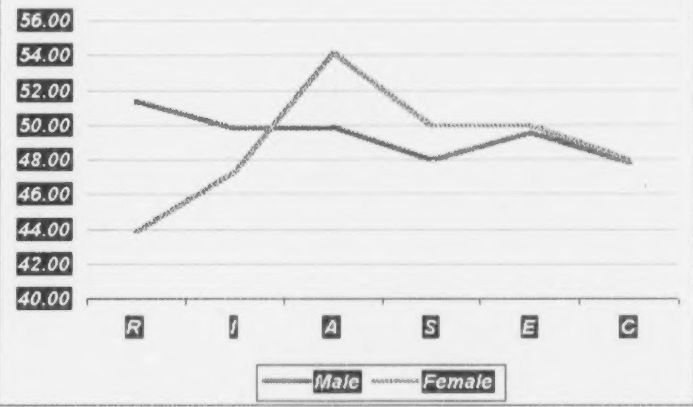
Cody S. Ding
Arizona State University

A great many current investigations, either in psychology or in education, deals with profiles of test scores. The terminology "profile" has been widely used in education settings to indicate a student's performance on a set of test scores such as reading, math, writing, and critical thinking skills. It is not uncommon that students receive test report with their score profiles, representing the strength and weakness in their performance on different tests. Due to this common practice in education, the profile analysis was sometimes considered by education practitioners as simply depicting test scores. In addition, a variety of exploratory techniques have been used to identify profile patterns in a set of data. These methods includes various cluster analytic approaches (e.g., Konald, Glutting, McDermott, Krush, & Watkins, 1999), configural frequency analysis (e.g., Stanton & Reynolds, 2001), and model profile analysis (e.g., Moses & Pritchard, 1995), which is a hybrid of cluster analysis and Cattell's (1967) Q-factor analysis. With the wide use of profile analysis, researchers and practitioners need to be aware of issues on the topic and how it should be dealt with. The goal of this paper is, thus, to introduce the problem and how it can be dealt with appropriately and provide an example of how profile analysis can be done, especially based on multidimensional scaling.

How does profile analysis work?

The term "profile" comes from the practice in applied work in which scores on a test battery are plotted in terms of graph or profile. Figure 1 shows an example of profiles for males and females on six variables of Strong Vocational Interest (Strong, 1955): Realistic, Investigative, Artistic, Social, Enterprising, and conventional.

Figure 1: Strong Vocational Interest Mean Profiles for Males and Females



As shown in Figure 1, the profile provides three types of information for any person or group: level, dispersion, and the shape. The profile level is defined as an unweighted average of the scores in the profile, that is, the mean score over the six interest variables. In Figure 1, we would obtain level by adding scores on the six interest variables and dividing by 6 for males and females, respectively. It seems that the level for males (level = 49.41) is higher than that for females (level = 48.91).

The profile dispersion is defined as how much each score in the profile deviates from the mean. A measure of the dispersion is the standard deviation of scores for each person or group. In Figure 1, we would compute dispersion for males and females by subtracting the scores on each of the six interest variables from each group's level. Whereas it is possible to make a direct interpretation of level, it is difficult to do so for dispersion because the profile dispersions for people in general depend on the correlations among variables in the profile. According to Nunnally (1978), the most sensible way to interpret the dispersion of scores is to compare the dispersions of scores for two people or groups. In Figure 1, profile for females has a much larger dispersion (dispersion = 3.43) than that for males (dispersion = 1.33) on the six interest variables. Thus, it may be concluded that females' interests are more variable than those of males.

The last information provided by profile is the profile shape. The profile shape is defined as the "ups" and "downs" in the profile and can be determined by the rank-order of scores. The high and low points in the profile indicate the salient characteristics of the person or group who resembles the profile. It should be noted that the actual appearance of a particular profile depends on the way variables are listed. Since it is arbitrary which variable is listed in which position, the physical appearance of the profile can be arbitrarily changed without impacting level, dispersion, or shape. Thus in Figure 1, females have the highest score on Artistic and the lowest score on Realistic, whereas males have the highest score on Realistic and lowest score on Convention.

The profiles in Figure 1 are the simplest way of profile analysis in that all that involved is to plot a set of scores of the variables for any given person or group, as we often do for educational testing data. Profile analysis is, however, a generic term for all methods concerning groupings of persons. There are two sets of research problems underlying the use of profile analysis. One set of problems in profile analysis is that group memberships are known in advance of the analysis. In this case, the purpose of the profile analysis is to distinguish the groups from one another on the basis of variables in the data matrix. One such a problem, for example, could be start with groups of males and females in a population on vocational interests, as we have done in Figure 1. The purpose of analysis would be to distinguish these two groups in terms of variables used to measure the traits. This type of problems relates to multivariate analysis of variance (MANOVA) or discriminate analysis, which is concerned with a priori groupings of people. Thus in profile analysis via MANOVA, one simultaneously test hypotheses of a priori group differences on a set of variables; on conversely, one test hypotheses, in discriminate analysis, of differentiation of a priori groups with a set of variables and forms linear combinations of those variables that will most effectively differentiate in that regard.

The second type of problems in profile analysis occurs when group membership of people are not stated in advance of the analysis; the purpose of analysis is, thus, to classify people according to their scores on set of variables. Accordingly, while profile analysis via MANOVA or discriminate analysis is focused on the hypothesis testing about the extent to which a priori groups hang together, the clustering profiles is focused on discovering groups of people that hang together. However, it is important to note that the problem of test for significance of difference between groups on a set of variables considered simultaneously should not be confused with the major problem of profile analysis that we are attempting to deal with here, namely, classify people via proper methods. This is the topic that we now turn to.

Profile analysis via Multidimensional Scaling approach (PAMS) model

As we mentioned above, there is a variety of exploratory techniques that have been used to identify profile patterns. This paper outlines an exploratory multidimensional scaling based approach to identifying the major profile patterns in the data, called Profile Analysis via Multidimensional Scaling (PAMS) that was originated by Davison (Davison, 1994). The Multidimensional Scaling is not a new technique but the profile interpretation of dimension is new. In comparison with other approaches of profile analysis (e.g., cluster analysis and configural frequency analysis), there are four major distinct features of the PAMS model. First, the model includes the Q-factor model (Cattell, 1967) as a special case. Furthermore, unlike techniques based on the Q-factor approach, the PAMS model can readily be applied to samples of any size. In fact, in PAMS analysis, the sample size consists of the number of variables, and to use that as N in the usual formulas forces one to consider random samples of variables in a domain.

Second, the PAMS model is designed for the study of latent "person." If one has theories concerning "types" among people, as opposed to "factors" among variables in factor analysis, one should use the PAMS model, which studies clusters of people and each cluster is a hypothetical "prototypical person." The prototypical person is defined in terms of "his/her" complete set of responses to the variables involved. Third, cluster analysis is very similar to factor analysis and it is designed to classify variables into "clusters" and it has been used to describe people in terms of discrete groupings. On the other hand, the PAMS model describes people in terms of continuous person profile indices that specify to what extent people are mixtures of the various types. That is, people are assumed to vary around each of several "prototypical" persons. Fourth, the person profile indices can be used either as predictors or criterion variables in subsequent analysis to study the relationships between individual profile patterns and other variables such as treatment outcomes. There is no such counterpart in other profile analysis technique procedures as far as the author knows.

I now turn to a description of the PAMS model and the analysis based on that model.

A full PAMS analysis involves two parts: (1) estimation of latent or prototypical profiles in a population, and (2) estimation of person profile index. Each person profile index quantifies the degree of correspondence between the observed profile of that individual and one of the latent or prototypical profiles. In this paper I will focus on the first step, estimation of latent or prototypical profiles in a population.

In classical MDS representations of structure, each data point is represented in a Euclidean space of continuous dimensions, that is:

$$m_{sv} = \sum_k w_{sk} x_{vk} + c_s + e_{vs} \quad (1)$$

Where m_{sv} is an observed score of subject s on variable v, c_s is a level parameter, and e_{vs} is an error term representing measurement error and systematic deviations from the model. Each term in the sum on the right side of Equation 1 refers to a latent profile pattern K and this latent profile pattern corresponds to a multidimensional scaling dimension. The x_{vk}, the scale value parameter, equals the scores of variables in latent profile K. The w_{sk}, person profile index, is a measure of profile match that indexes the degree of match between the observed profile of subject s and the latent profile K. In an analysis based on the model of equation (1), the goal is to estimate the number of dimensions K , the scale value parameters, x_{vk}, along each dimension, the person profile index w_{sk}, and a measure of fit.

To estimate those parameters, one needs a set of assumptions and restrictions that uniquely identify the parameters (See Davison, 1996). One of the important assumptions is ipsative, which states that the mean of the scores in each latent profile equals zero, that is, $\Sigma_v x_{vk} = 0.0$ for all K. Consequently, latent profiles will reproduce observed profile patterns (i.e., scatter plus shape), but not the level of the observed profiles which is reproduced by the level parameters. Under those assumptions, one can perform the parameter estimation based on the Squared Euclidean distance that satisfies the fundamental assumptions of common MDS analyses. One can, hence, estimate the scale value parameters in the PAMS model (i.e., Equation 1) by computing the squared Euclidean distance matrix for all possible pairs of variables and submitting that matrix to an MDS analysis that can be found in major statistical packages. The MDS analysis should produce one dimension for each latent profile K, along which the scale value, x_{vk}, for variable v is our estimate of the score for that variable in latent profile. The number of dimensions or latent profiles may be determined based on theory, previous research, or statistical methods. One such method is to use badness-of-fit index of .05 or less. One can choose a given dimensional solution based on which dimensional solution produces the smallest badness of fit index by the MDS analysis.

The PAMS model leads to the interpretation of MDS dimensions as latent profiles. Each MDS dimension K represents a prototype individual, that is, each MDS dimension represents a group of individuals with the similar characteristics. How does one interpret the level parameter c_s, scale value x_{vk}, and person profile index w_{sk}? The parameter c_s equals the mean score in row; that is, it indexes the overall elevation of subject's profile, and it is therefore called level parameter. As mentioned earlier, Q-factor model (Cattell, 1967) is a special case of the PAMS model in which the observed data are standardized across subjects and therefore this level parameter, c_s, drops from the model because it equals zero for every subject s.

The set of scale values x_{vk} along each dimension represents a profile of scores for that prototypical person k on measure v. They indicate deviations from the profile level. The person profile index w_{sk} indicates how well a person resembles the prototypical or latent profile. The product of the person profile index, w_{sk}, and the scale values, x_{vk}, forms the profile patterns that provide addition information that is beyond and above that provided by the profile level (i.e., the average score). In contrast, in the profile analysis of priori defined group of individuals via analysis of variance, the information provided by the analysis is only based on profile level c_s of equation 1, as we illustrated in Figure 1.

An example

To illustrate the profile analysis via Multidimensional Scaling (PAMS) approach, I used the same six variables of Strong Vocational Interest as in Figure 1 to identify "types" of people who have similar configurations of interests. Analysis was run for the total sample of 344 subjects since the emphasis here was on the results for the total sample. The scale value parameters in the model were estimated using the most common, nonmetric MDS analyses found in SAS (or other most major computer packages). The 2 dimensional solution was obtained using a non-metric scaling MDS procedure in SAS. For an illustrative purpose, below I included data for 10 subjects, the brief descriptions of the variables (the full discussion of the original terms can be found in Holland's 1963 paper) , and the SAS program codes to perform the analysis.

```
Obs RTHEME ITHEME ATHEME STHEME ETHEME CTHEME
 1  45  39  39  38  56  51
 2  47  51  56  53  48  47
 3  65  67  51  39  54  55
 4  45  54  43  45  45  45
 5  52  47  51  49  49  54
 6  64  52  53  32  58  45
 7  54  57  29  45  56  62
 8  52  57  39  57  59  54
 9  60  57  55  64  65  54
10  49  54  57  47  48  48
```

RTHEM = Realistic, indicating people who enjoy activities requiring physical strength, aggressive action, motor coordination and skill.

ITHEME = Investigative, indicating task-oriented people who generally prefer to think through rather than act out problems.

ATHEME = Artistic, indicating people who prefer indirect relations with others, that is, dealing with environmental problems through self expression in artistic media.

STHEME = Social, indicating people who prefer teaching role, which may reflect a desire for attention and socialization in a structured setting.

ETHEME = Enterprising, indicating people who prefer to use their verbal skills in situations which provide opportunities for dominating, selling, or leading others.

CTHEME = Conventional, indicating people who prefer structured verbal and numerical activities, and subordinate role.

```
OPTIONS nocenter nodate;

* Read in the data;
DATA one;
INFILE 'a:interest';

* Create dissimilarity matrix between variables;
PROC transpose DATA=one OUT=out ;
VAR rtheme--ctheme;
RUN;

%INCLUDE 'C:\program files\SAS institute\sas\v8\STAT\SAMPLE\xmacro.SAS';
%INCLUDE 'C:\program files\SAS institute\sas\v8\STAT\SAMPLE\distnew.SAS';
%INCLUDE 'C:\program files\SAS institute\sas\v8\STAT\SAMPLE\stdize.SAS';
```

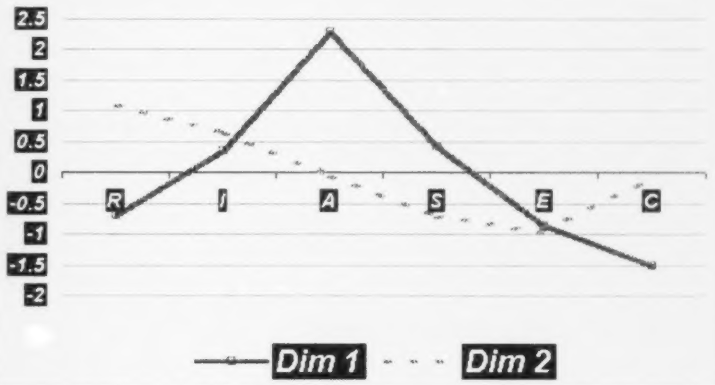
```
%DISTANCE ( DATA=out,METHOD=EUCLID, id=_name_, OUT=two );
RUN;

* Perform MDS analysis.
PROC MDS
DATA=two COEF=i DEC=2 DIM=2 FORMULA=1 LEVEL=ordinal FIT=2
CONDITION=matrix SHAPE=square NOP PFINAL OUT=coord;
RUN;
```

Table 1 shows the estimated scale values for these two dimensions and these scale values were plotted in Figure 2 as the latent profiles of the six interest variables.

Table 1: Estimated 2-Dimenional Scale Values of the Six Interest Variables

	Dim 1	Dim 2
RTHEME	-0.70	1.09
ITHEME	0.36	0.65
ATHEME	2.28	-0.04
STHEME	0.42	-0.69
ETHEME	-0.85	-0.94
CTHEME	-1.51	-0.07

Figure 2: Strong Vocational Interest Profiles



As I mentioned above, the actual appearance of a particular profile depends on the way variables are listed and since it is arbitrary which variable is listed in which position, the physical appearance of the profile can be arbitrarily changed without impacting profile level and profile patterns. Because of this reason, some researchers call high points in a latent profile a profile and low points in the latent profile a mirror image of the profile (Davison, 1994).

As shown in Figure 2, Dimension 1 has a high score on Artistic on the positive end and Convention/Enterprising on the negative end. This dimension looks like

Prediger's (1982) Data vs Ideas Dimension. According to theory, the Data profile should be marked by the Conventional/Enterprising scales and the Ideas profile should be marked by the Artistic/Investigative scales, although in our profile the Investigative scale does not appear to be that salient as theory would lead one to expect. Thus, those whose interest profile resembles the Artistic vs. Convention profile shapes tend to be either more interested in Artistic than in Convention or more in Convention than in Artistic.

On the other hand, Dimension 2 has a high score on Realistic on the positive end and high scores on Social and Enterprising at the negative end. This dimension represents, according to theory, People vs. Things latent profiles (Prediger, 1982). People whose vocational interest profile resembles this profile shape tend to either involve in more realistic-type interest than social-type interest or involve in more social-type interest than realistic-type interest. That is, a subject resembles People profile would rather work with people than things and a subject resembles Things profile would rather work with things than people. Taken together, Dimensions 1 and 2 constitute a spatial representation of the major profile patterns in the data matrix as recovered by MDS.

Conclusion

In this paper I examined several different approaches of profile analysis. Particularly, I illustrated profile analysis by simply depicting a set of test scores and that by PAMS model. Proper choice of a method for a specific investigation requires knowledge of the assumptions, limitations, and information utilized in the methods of measuring profile. It is important not to confuse the clustering of person in the PAMS approach with the group comparison on a set of scores in multivariate analysis of variance approach (MANOVA), as we did in the first example. It is also important to distinguish analyzing variables as in factor analysis from analyzing people as we did in here. In this latter approach, one is concerned with identifying prototypical persons in a population rather than identifying constructs from a set of variables. The PAMS model has advantages of being easily applied to samples of any size, classifying person on a continuum scale, and using person profile index for further hypothesis studies, but there are some caveats that need to be noted. First, the determination of number of dimensions or latent profiles is based on interpretability, reproducibility, and the fit statistics. Second, the interpretation of significance of the scale values is somewhat arbitrary. There is no objective criteria for decision making regarding which scale values are salient. Some researchers (e.g., Kim and Davison, 2001) used bootstrapping method to estimate standard error of scale values and statistical significance of these scale values was used for making decisions. Third, it is not well known as to what degree the latent profiles recovered by MDS solutions can be generalized across populations. Based on the limited research so far, it seems that different latent profiles may be recovered in two different populations such as in female population and male population, although the same latent profile solutions are expected to emerge on different data sets in the same population. More researches are needed in these areas.

Note

A second dataset and corresponding SAS control cards can be found at http://ericae.net/pare/htm/v7n16/profileanalysis.htm

References

Cattell, R. B. (1967). The three basic factor analysis research designs: Their interrelations and derivatives. In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment*. New York: McGraw-Hill.

Davison, M. L. (1994). Multidimensional scaling models of personality responding. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality*. New York: Springer.

Davison, M. L. (1996). Addendum to "Multidimensional scaling and factor models of test and item responses." Unpublished report, Department of Educational Psychology, University of Minnesota.

Davison, M. L., Blake, R., & Sackett, S. (August, 1997). Relationships between personality and interest profiles: A canonical regression approach. Paper presented at the annual meeting of the American Psychological Association, Washington D.C.

Holland, J. L. (1963). Explorations of a theory of vocational choice and achievement: II. A four-year prediction study. *Psychological Reports, 12*, 547-594.

Kim, S. & Davison, M. L. (2001). Bootstrapping for estimating standard errors in multidimensional scaling. Unpublished paper presented to the American Educational Research Association, Seattle, WA.

Konald, T. R., Glutting, J.J., McDermott, P.A., Kush, J. C., & Watkins, M.M. (1999). Structure and diagnostic benefits of a normative subtest taxonomy developed from the WIS-II standardization sample. *Journal of School Psychology, 37*, 29-48.

Moses, J. A., Jr. & Pritchard, D. A. (1995). Modal profiles for the Wechsler Adult Intelligence Scale-Revised. *Archives of Clinical Neuropsychology, 11*, 61-68.

Nunnally, J. C. (1978). *Psychometric theory* (2nd. ed). New York: McGraw-Hill.

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations. *Journal of Vocational*

Behavior, 21, 259-287.

Stanton, H. C. & Reynolds, C. R. (2001). Configural frequency analysis as a method of determining Wechsler profile types. *School Psychology Quarterly, 15,* 434-448.

Strong, E. K. Jr. (1955). *Vocational interests 18 years after college.* Minneapolis, MN: University of Minnesota Press.

Tabachnick, B. G. & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: HarperCollins Publisher.

Author Notes

Correspondence relating to this article can be addressed to Cody Ding, 301 Payne Hall, Arizona State University, Tempe, AZ 85287, or via email at cody.ding@asu.edu. The author would like to express his acknowledgments and appreciation to Mark Davison who provided a book and the articles from which I have drawn many of my ideas and arguments relating to this paper.

Descriptors: Profile analysis; scaling

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Stemler, Steve (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17). Available online: http://ericae.net/pare/getvn.asp?v=7&n=17. This paper has been viewed 1882 times since 6/7/01.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Stemler, Steve

An Overview of Content Analysis

Steve Stemler
Boston College

Content analysis has been defined as a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding (Berelson, 1952; GAO, 1996; Krippendorff, 1980; and Weber, 1990). Holsti (1969) offers a broad definition of content analysis as, "any technique for making inferences by objectively and systematically identifying specified characteristics of messages" (p. 14). Under Holsti's definition, the technique of content analysis is not restricted to the domain of textual analysis, but may be applied to other areas such as coding student drawings (Wheelock, Haney, & Bebell, 2000), or coding of actions observed in videotaped studies (Stigler, Gonzales, Kawanaka, Knoll, & Serrano, 1999). In order to allow for replication, however, the technique can only be applied to data that are durable in nature.

Content analysis enables researchers to sift through large volumes of data with relative ease in a systematic fashion (GAO, 1996). It can be a useful technique for allowing us to discover and describe the focus of individual, group, institutional, or social attention (Weber, 1990). It also allows inferences to be made which can then be corroborated using other methods of data collection. Krippendorff (1980) notes that "[m]uch content analysis research is motivated by the search for techniques to infer from symbolic data what would be either too costly, no longer possible, or too obtrusive by the use of other techniques" (p. 51).

Practical Applications of Content Analysis

Content analysis can be a powerful tool for determining authorship. For instance, one technique for determining authorship is to compile a list of suspected authors, examine their prior writings, and correlate the frequency of nouns or function words to help build a case for the probability of each person's authorship of the data of interest. Mosteller and Wallace (1964) used Bayesian techniques based on word frequency to show that Madison was indeed the author of the Federalist papers; recently, Foster (1996) used a more holistic approach in order to determine the identity of the anonymous author of the 1992 book *Primary Colors*.

Content analysis is also useful for examining trends and patterns in documents. For example, Stemler and Bebell (1998) conducted a content analysis of school mission statements to make some inferences about what schools hold as their primary reasons for existence. One of the major research questions was whether the criteria being used to measure program effectiveness (e.g., academic test scores) were aligned with the overall program objectives or reason for existence.

Additionally, content analysis provides an empirical basis for monitoring shifts in public opinion. Data collected from the mission statements project in the late 1990s can be objectively compared to data collected at some point in the future to determine if policy changes related to standards-based reform have manifested themselves in school mission statements.

Conducting a Content Analysis

According to Krippendorff (1980), six questions must be addressed in every content analysis:

1) Which data are analyzed?
2) How are they defined?
3) What is the population from which they are drawn?
4) What is the context relative to which the data are analyzed?
5) What are the boundaries of the analysis?
6) What is the target of the inferences?

At least three problems can occur when documents are being assembled for content analysis. First, when a substantial number of documents from the population are missing, the content analysis must be abandoned. Second, inappropriate records (e.g., ones that do not match the definition of the document required for analysis) should be discarded, but a record should be kept of the reasons. Finally, some documents might match the requirements for analysis but just be uncodable because they contain missing passages or ambiguous content (GAO, 1996).

Analyzing the Data

Perhaps the most common notion in qualitative research is that a content analysis simply means doing a word-frequency count. The assumption made is that the words that are mentioned most often are the words that reflect the greatest concerns. While this may be true in some cases, there are several counterpoints to consider when using simple word frequency counts to make inferences about matters of importance.

One thing to consider is that synonyms may be used for stylistic reasons throughout a document and thus may lead the researchers to underestimate the importance of a concept (Weber, 1990). Also bear in mind that each word may not represent a category equally well. Unfortunately, there are no well-developed weighting procedures, so for now, using word counts requires the researcher to be aware of this limitation. Furthermore, Weber reminds us that, "not all issues are equally difficult to raise. In contemporary America it may well be easier for political parties to address economic issues such as trade and deficits than the history and current plight of Native American living precariously on reservations" (1990, p. 73). Finally, in performing word frequency counts, one should bear in mind that some words may have multiple meanings. For instance the word "state" could mean a political body, a situation, or a verb meaning "to speak."

A good rule of thumb to follow in the analysis is to use word frequency counts to identify words of potential interest, and then to use a Key Word In Context (KWIC) search to test for the consistency of usage of words. Most qualitative research software (e.g., NUD*IST, HyperRESEARCH) allows the researcher to pull up the sentence in which that word was used so that he or she can see the word in some context. This procedure will help to strengthen the validity of the inferences that are being made from the data. Certain software packages (e.g., the revised General Inquirer) are able to incorporate artificial intelligence systems that can differentiate between the same word used with two different meanings based on context (Rosenberg, Schnurr, & Oxman, 1990). There are a number of different software packages available that will help to facilitate content analyses (see further information at the end of this paper).

Content analysis extends far beyond simple word counts, however. What makes the technique particularly rich and meaningful is its reliance on coding and categorizing of the data. The basics of categorizing can be summed up in these quotes: "A category is a group of words with similar meaning or connotations" (Weber, 1990, p. 37). "Categories must be mutually exclusive and exhaustive" (GAO, 1996, p. 20). Mutually exclusive categories exist when no unit falls between two data points, and each unit is represented by only one data point. The requirement of exhaustive categories is met when the data language represents all recording units without exception.

Emergent vs. *a priori* coding. There are two approaches to coding data that operate with slightly different rules. With *emergent coding*, categories are established following some preliminary examination of the data. The steps to follow are outlined in Haney, Russell, Gulek, & Fierros (1998) and will be summarized here. First, two people independently review the material and come up with a set of features that form a checklist. Second, the researchers compare notes and reconcile any differences that show up on their initial checklists. Third, the researchers use a consolidated checklist to independently apply coding. Fourth, the researchers check the reliability of the coding (a 95% agreement is suggested; .8 for Cohen's kappa). If the level of reliability is not acceptable, then the researchers repeat the previous steps. Once the reliability has been established, the coding is applied on a large-scale basis. The final stage is a periodic quality control check.

When dealing with *a priori* coding, the categories are established prior to the analysis based upon some theory. Professional colleagues agree on the categories, and the coding is applied to the data. Revisions are made as necessary, and the categories are tightened up to the point that maximizes mutual exclusivity and exhaustiveness (Weber, 1990).

Coding units. There are several different ways of defining coding units. The first way is to define them physically in terms of their natural or intuitive borders. For instance, newspaper articles, letters, or poems all have natural boundaries. The second way to define the recording units syntactically, that is, to use the separations created by the author, such as words, sentences, or paragraphs. A third way to define them is to use referential units. Referential units refer to the way a unit is represented. For example a paper might refer to George W. Bush as "President Bush," "the 43rd president of the United States," or "W." Referential units are useful when we are interested in making inferences about attitudes, values, or preferences. A fourth method of defining coding units is by using propositional units. Propositional units are perhaps the most complex method of defining coding units because they work by breaking down the text in order to examine underlying assumptions. For example, in a sentence that would read, "Investors took another hit as the stock market continued its descent," we would break it down to: The stock market has been performing poorly recently/Investors have been losing money (Krippendorff, 1980).

Typically, three kinds of units are employed in content analysis: sampling units, context units, and recording units.

- *Sampling units* will vary depending on how the researcher makes meaning; they could be words, sentences, or paragraphs. In the mission statements project, the sampling unit was the mission statement.

- *Context units* neither need be independent or separately describable. They may overlap and contain many recording units. Context units do, however, set physical limits on what kind of data you are trying to record. In the mission statements project, the context units are sentences. This was an arbitrary decision, and the context unit could just as easily have been paragraphs or entire statements of purpose.

- *Recording units*, by contrast, are rarely defined in terms of physical boundaries. In the mission statements project, the recording unit was the idea (s) regarding the purpose of school found in the mission statements (e.g., develop responsible citizens or promote student self-worth). Thus a sentence that reads "The mission of Jason Lee school is to enhance students' social skills, develop responsible citizens, and foster emotional growth" could be coded in three separate recording units, with each idea belonging to only one category (Krippendorff, 1980).

Reliability. Weber (1990) notes: "To make valid inferences from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way" (p. 12). As Weber further notes, "reliability problems usually grow out of the ambiguity of word meanings, category definitions, or other coding rules" (p. 15). Yet, it is important to recognize that the people who have developed the coding scheme have often been working so closely on the project that they have established shared and hidden meanings of the coding. The obvious result is that the reliability coefficient they report is artificially inflated (Krippendorff, 1980). In order to avoid this, one of the most critical steps in content analysis involves developing a set of explicit recording instructions. These instructions then allow outside coders to be trained until reliability requirements are met.

Reliability may be discussed in the following terms:

- *Stability*, or intra-rater reliability. Can the same coder get the same results try after try?
- *Reproducibility*, or inter-rater reliability. Do coding schemes lead to the same text being coded in the same category by different people?

One way to measure reliability is to measure the percent of agreement between raters. This involves simply adding up the number of cases that were coded the same way by the two raters and dividing by the total number of cases. The problem with a percent agreement approach, however, is that it does not account for the fact that raters are expected to agree with each other a certain percentage of the time simply based on chance (Cohen, 1960). In order to combat this shortfall, reliability may be calculated by using Cohen's Kappa, which approaches 1 as coding is perfectly reliable and goes to 0 when there is no agreement other than what would be expected by chance (Haney et al., 1998). Kappa is computed as:

$$\kappa = \frac{P_A - P_c}{1 - P_c}$$

where:

P_A = proportion of units on which the raters agree
P_c = the proportion of units for which agreement is expected by chance.

Table 1 – Example Agreement Matrix

		Rater 1			Marginal Totals
		Academic	Emotional	Physical	
Rater 2	Academic	.42 (.29)*	.10 (.21)	.05 (.07)	**.57**
	Emotional	.07 (.18)	.25 (.18)	.03 (.05)	**.35**
	Physical	.01 (.04)	.02 (.03)	.05 (.01)	**.08**
		.50	**.37**	**.13**	**1.00**

*Values in parentheses represent the expected proportions on the basis of chance associations, i.e. the joint probabilities of the marginal proportions.

Given the data in Table 1, a percent agreement calculation can be derived by summing the values found in the diagonals (i.e., the proportion of times that the two raters agreed):

$$P_A = .42 + .25 + .05 = .72$$

By multiplying the marginal values, we can arrive at an expected proportion for each cell (reported in parentheses in the table). Summing the product of the marginal values in the diagonal we find that on the basis of chance alone, we expect an observed agreement value of:

$$P_C = .29 + .18 + .01 = .48$$

Kappa provides an adjustment for this chance agreement factor. Thus, for the data in Table 1, kappa would be calculated as:

$$\kappa = \frac{.72 - .48}{1 - .48} = .462$$

In practice, this value may be interpreted as the proportion of agreement between raters after accounting for chance (Cohen, 1960). Crocker & Algina (1986) point out that a value of $\kappa = 0$ does not mean that the coding decisions are so inconsistent as to be worthless, rather, $\kappa = 0$ may be interpreted to mean that the decisions are no more consistent than we would expect based on chance, and a negative value of kappa reveals that the observed agreement is worse than expected on the basis of chance alone. "In his methodological note on kappa in *Psychological Reports*, Kvalseth (1989) suggests that a kappa coefficient of 0.61 represents reasonably good overall agreement." (Wheelock et al., 2000). In addition, Landis & Koch (1977, p.165) have suggested the following benchmarks for interpreting kappa:

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00– 0.20	Slight
0.21– 0.40	Fair
0.41– 0.60	Moderate
0.61– 0.80	Substantial
0.81– 1.00	Almost Perfect

Cohen (1960) notes that there are three assumptions to attend to in using this measure. First, the units of analysis must be independent. For example, each mission statement that was coded was independent of all others. This assumption would be violated if in attempting to look at school mission statements, the same district level mission statement was coded for two different schools within the same district in the sample.

Second, the categories of the nominal scale must be independent, mutually exclusive, and exhaustive. Suppose the goal of an analysis was to code the kinds of courses offered at a particular school. Now suppose that a coding scheme was devised that had five classification groups: mathematics, science, literature, biology, and calculus. The categories on the scale would no longer be independent or mutually exclusive because whenever a biology course is encountered it also would be coded as a science course. Similarly, a calculus would always be coded into two categories as well, calculus and mathematics. Finally, the five categories listed are not mutually exhaustive of all of the different types of courses that are likely to be offered at a school. For example, a foreign language course could not be adequately described by any of the five categories.

The third assumption when using kappa is that the raters are operating independently. In other words, two raters should not be working together to come to a consensus about what rating they will give.

Validity. It is important to recognize that a methodology is always employed in the service of a research question. As such, validation of the inferences made on the basis of data from one analytic approach demands the use of multiple sources of information. If at all possible, the researcher should try to have some sort of validation study built into the design. In qualitative research, validation takes the form of triangulation. Triangulation lends credibility to the findings by incorporating multiple sources of data, methods, investigators, or theories (Erlandson, Harris, Skipper, & Allen, 1993).

For example, in the mission statements project, the research question was aimed at discovering the purpose of school from the perspective of the institution. In order to cross-validate the findings from a content analysis, schoolmasters and those making hiring decisions could be interviewed about the emphasis placed upon the school's mission statement when hiring prospective teachers to get a sense of the extent to which a school's values are truly reflected by mission statements. Another way to validate the inferences would be to survey students and teachers regarding the mission statement to see the level of awareness of the aims of the school. A third option would be to take a look at the degree to which the ideals mentioned in the mission statement are being implemented in the classrooms.

Shapiro & Markoff (1997) assert that content analysis itself is only valid and meaningful to the extent that the results are related to other measures. From this perspective, an exploration of the relationship between average student achievement on cognitive measures and the emphasis on cognitive outcomes stated across school mission statements would enhance the validity of the findings. For further discussions related to the validity of content analysis see Roberts (1997), Erlandson et al. (1993), and Denzin & Lincoln (1994).

Conclusion

When used properly, content analysis is a powerful data reduction technique. Its major benefit comes from the fact that it is a systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding. It has the attractive features of being unobtrusive, and being useful in dealing with large volumes of data. The technique of content analysis extends far beyond simple word frequency counts. Many limitations of word counts have been discussed and methods of extending content analysis to enhance the utility of the analysis have been addressed. Two fatal flaws that destroy the utility of a content analysis are faulty definitions of categories and non-mutually exclusive and exhaustive categories.

Further information

For links, articles, software and resources see

`http://writing.colostate.edu/references/research/content/`

`http://www.gsu.edu/~wwwcom/.`

References

Berelson, B. (1952). *Content Analysis in Communication Research*. Glencoe, Ill. Free Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, pp. 37– 46.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Harcourt Brace Jovanovich.

Denzin, N.K., & Lincoln, Y.S. (Eds.). (1994). *Handbook of Qualitative Research*. Thousand Oaks, CA: Sage Publications.

Erlandson, D.A., Harris, E.L., Skipper, B.L., & Allen, S.D. (1993). *Doing Naturalistic Inquiry: A Guide to Methods*. Newbury Park, CA: Sage Publications.

Foster, D. (1996, February 26). Primary culprit. *New York*, 50– 57.

Haney, W., Russell, M., Gulek, C., and Fierros, E. (Jan-Feb, 1998). Drawing on education: Using student drawings to promote middle school improvement. *Schools in the Middle, 7(3)*, 38– 43.

Holsti, O.R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Newbury Park, CA: Sage.

Kvalseth, T. O. (1989). Note on Cohen's kappa. *Psychological reports*, 65, 223– 26.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, pp. 159– 174.

Mosteller, F. and D.L. Wallace (1964). *Inference and Disputed Authorship: The Federalist*. Reading, Massachusetts: Addison-Wesley.

Nitko, A.J. (1983). *Educational Tests and Measurement: An Introduction*. New York, NY: Harcourt Brace Jovanovich.

Roberts, C.W. (Ed.) (1997). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rosenberg, S.D., Schnurr, P.P., & Oxman, T.E. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment, 54* (1 & 2), 298– 310.

Shapiro, G., & Markoff, J. (1997). 'A Matter of Definition' in C.W. Roberts (Ed.). *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stemler, S., and Bebell, D. (1998). *An Empirical Approach to Understanding and Analyzing the Mission Statements of Selected Educational Institutions*. Paper presented at the annual meeting of the New England Educational Research

Organization. Portsmouth, New Hampshire. Available: ERIC Doc No. ED 442 202.

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S. & Serrano, A. (1999). *The TIMSS Videotape Classroom Study: Methods and Findings from an Exploratory Research Project on Eighth-Grade Mathematics Instruction in Germany, Japan, and the United States.* U.S. Department of Education National Center for Educational Statistics: NCES 99-074. Washington, D.C.: Government Printing Office.

U.S. General Accounting Office (1996). *Content Analysis: A Methodology for Structuring and Analyzing Written Material.* GAO/PEMD-10.3.1. Washington, D.C. (This book can be ordered free from the GAO).

Weber, R. P. (1990). *Basic Content Analysis*, 2nd ed. Newbury Park, CA.

Wheelock, A., Haney, W., & Bebell, D. (2000). What can student drawings tell us about high-stakes testing in Massachusetts? *TCRecord.org*. Available: http://www.tcrecord.org/Content.asp?ContentID=10634.

Please address all correspondence regarding this article to:

Steve Stemler, Ph.D.
Boston College
TIMSS International Study Center
140 Commonwealth Avenue
Chestnut Hill, MA 02467

E-Mail stemler@bc.edu
http://www2.bc.edu/~stemler

Descriptors: Content analysis; NUD*IST; Research Methods; Qualitative Analysis

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Simon, Marielle & Renée Forgette-Giroux (2001). A rubric for scoring postsecondary academic skills. *Practical Assessment, Research & Evaluation*, 7(18). Available online: http://ericae.net/pare/getvn.asp?v=7&n=18. This paper has been viewed 1115 times since 6/17/01.

- Find similar papers in
  - ERICAE Full Text Library
  - *Pract Assess, Res & Eval*
  - ERIC RIE & CIJE 1990-
  - ERIC On-Demand Docs
  - ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  - Simon, Marielle
  - Renée Forgette-Giroux

A RUBRIC FOR SCORING POSTSECONDARY ACADEMIC SKILLS

Marielle Simon & Renée Forgette-Giroux,
Faculty of Education, University of Ottawa

Today's assessment of postsecondary academic skills must take into account their comprehensive nature and their multiple facets (Biggs, 1995; Sadler, 1989). In this regard, the use of rubric is more likely to provide qualitative, meaningful, and stable appraisals than are traditional scoring methods. The stability of assessment results, however, rests on the scale's ability to lead to a common and uniform interpretation of student performance. The assessment of postsecondary academic skills on the basis of such a scale offers several advantages. First, it presents a continuum of performance levels, defined in terms of selected criteria, towards to full attainment or development of the targeted skills. Second, it provides qualitative information regarding the observed performance in relation to a desired one. Third, its application, at regular intervals, tracks the student's progress of his or her skill mastery. Finally, the choice of rather broad universal criteria extends the application to several contexts.

Despite its merits, however, the use of a generic descriptive scale at the postsecondary level is relatively recent and some difficulties need to be addressed. This paper has three objectives:

a. to present the nature of a generic rubric used to assess postsecondary academic skills,

b. to describe a preliminary application in a university setting, and
c. to discuss observed related issues from a research point of view.

Nature of the rubric

The rubric for scoring academic skills is essentially qualitative and descriptive in nature and relies on criterion-referenced perspectives. It serves to appraise academic competencies such as the ability to critique, to produce scholarly work, to synthesize, and to apply newly acquired principles and concepts. It requires the use of criteria that best describe actual student products in a postsecondary setting. The criteria form the left-hand column of the two-way table format and the horizontal continuum contains headings indicating four increasing levels of performance towards competency mastery (Wiggins, 1998).

The use of the scale involves the acts of scoring, interpreting, and judging. (Forgette-Giroux, & Simon, 1998; Simon, & Forgette-Giroux, 2000). Scoring occurs when one identifies, within the scale, and for each criterion, the cell description that most closely matches the observed performance. The interpretation consists of locating the column that best describes the level of skill mastery. Judging means comparing the identified or observed performance level to a predetermined standard level.

Context of Application

The rubric discussed in this paper has evolved over the past five years but the latest, most generic version, was used within four graduate and two undergraduate level courses. Course enrollment varied from three to 30 students for a total of approximately 100 students. The courses were taught by the two authors, both experts in measurement and evaluation in education, and their topics related to research methodology or assessment. Given their theoretical nature, all courses were organized to assist students in their development and mastery of a single, carefully formulated academic skill, such as the ability to critically analyze a variety of research studies in education, to write a research proposal or report in education, and to assess student learning using current assessment methods and principles. Students were asked to assemble a portfolio that included scholarly works such as critiques, proposals, essays, manuscripts (Forgette-Giroux, et al., 1998). Practical assignments such as lesson plans, tests, performance assessments, were always accompanied by a structured critique. Students used the scale to self-assessed their portfolio for formative and summative purposes.

In this specific context, the performance levels, or anchors, are labeled as *good, very good, excellent*, and *exceptional*, to conform to the university approved grading scale. The five criteria are: Relevance, scope, accuracy, coherence, and depth. These criteria are commonly applied to scholarly writing by most manuscript review processes (NCME, PARE)[1], as are other attributes such as clarity, rigor, appeal, and strength of argument. The five criteria are also those found in the curriculum scoring rubrics mandated by the regional educational jurisdiction. The latter were to be learned by the training teachers at the undergraduate level and eventually used in their future teaching environment.

During the repeated application of the scale to the various university level courses, three concerns arose that have also been noticed elsewhere, and which continue to interest researchers. The following sections describe these difficulties and present their tentative treatment in this particular university setting.

Scale levels (anchors) identification

When the stages of development or mastery of the targeted skills are not empirically grounded, the initial identification of the scale levels is often arbitrarily determined. Also, when courses are given for the first time, the lack of student work samples further complicates the scale level identification process. Some researchers define scale levels and criteria in a *post hoc* fashion, such as was the case with the National Assessment of Educational Progress (Burstein, Koretz, Linn, Sugrue, Novak, Baker, & Lewis Harris, 1995/1996). The difficulty with this approach is that it is context specific and students cannot be made aware of these parameters prior to the assessment. An alternative procedure is to select work from the student at hand, that is typical of the upper levels of the scale or of the standard level. Wiggins (1998) suggests that, given clear parameters around the intended use of the rubric, those criteria that make the most sense are chosen with an understanding that they may be constantly adjusted based on exemplary performances. In the university context described here, the first version of the scale was developed around the expected student performance at the level of *excellence*. As the course progressed, performance exemplars of that level were identified, distributed among the students, and used to refine the scale.

Specificity of descriptors

For the scale to be generic enough to be applied in a variety of university courses, the descriptors need to refer to a spread of performances at each level. On the other hand, there is a risk that these statements may be too general and thus lead to inconsistent interpretation of the data. In the study reported here, the descriptors were formulated based on criteria associated with the development of valued academic skills that are relatively independent of the course contents. These skills tend to combine declarative and procedural knowledge with scholarly writing. The universality and pertinence of the selected criteria in terms of academic and practical perspectives extended the applicability of the descriptors to a variety of courses at both undergraduate and graduate levels and ensured student endorsement. In addition, the formative assessment at least once during the course, allowed the students and their professor to mediate scale interpretation in order to produce stable results. Despite its early stages of development, the scale yielded average percent agreements of 75 % between professor and student ratings.

Qualitative rating versus quantitative scoring

Student bodies and administrative pressures stress the attribution of a letter grade or a quantitative score to ratings obtained using the descriptive scale. In assigning a score, the rubric loses its ability to provide detailed and meaningful information about the quality of the performance as reflective of a specific level of skill mastery. Within the study context, the university administration required the presentation of a letter grade. Its scale equates *Exceptional* with the letter A+, *Excellent* with A/A-, *Very good* with B+/B and *Good* with C+. Throughout each course, assessment results were communicated to the students primarily using descriptive statements based on the rubric, but a final letter grade had to be assigned at the end of the course for official transcript purposes. It is interesting to note that, in adopting the scale for their own courses, colleagues typically experienced the need to quantify their assessment using complex algorithms, medians, modes, averages. In doing so, they easily lost track of the object of assessment. It would appear therefore, that the transition toward a purely qualitative approach within certain administrative constraints, takes repeated applications, discussion, and much self-reflection.

Discussion and Conclusion

The rubric was initially conceived as a substitute for the numerical scale that became obsolete and unstable in its traditional application, particularly when assessing complex skills through performance assessments. Its usefulness in higher education, therefore, largely depends on its ability to lead to meaningful and stable assessment results. Relevancy and consistency of results refer to validity and reliability issues. Among some of the design considerations put forward by Arter (1993) in the selection of good criteria when constructing rubrics for performance assessments, the most relevant to postsecondary contexts are (a) the need for universal attributes, (b) the means for assessing both holistically and analytically, and (c) the identification of the main components of the object of assessment. Moskal and Laydens (2000) have proposed practical ways to address these issues. They equate evidence related to content with the extent to which the rubric relates to the subject domain, and construct-related evidence to the conceptualization of a complex internal skill. Criterion-related evidence, meanwhile, serves to indicate how well the scoring criteria match those found in practice. Given this rubric's generic nature and the focus on the assessment of academic skills, primary attention must be given to the production of construct-related evidence. This was achieved by linking the scale's criteria, anchors, and descriptors to the nature of the skill addressed by the rubric and expressed in terms of a single learning objective.

Interrater and intrarater aspects of reliability were greatly improved by attaching the rubric to the course outline and by clarifying its various components and use early in the course, by enabling the students to access high quality exemplars, by providing regular qualitative feedback, by inviting the students to take part in mediation during formative assessments, and by requesting them to justify, in writing, their self-assessment based on specific references to their portfolio. It was important that this written rationale clearly support their perceived level of achievement. Written support of scoring decisions by the professor was also expected.

Given the exploratory nature of the study, many questions arise. However, five are of particular interest from both practical and research perspectives:

1. Would a scale based on a combination of the *post hoc* approach, of theoretical foundations of academic skills, and of samples of student work lead to increased validity?
2. Would a scale based on a combination of both qualitative and quantitative components lead to even greater consistency of results when assessing academic skills?
3. Is there an optimal class size to which this rubric can be applied most efficiently?
4. What type of teaching style is more likely to fit with the effective use of the scale?
5. Would the scale be as useful in courses focusing on content rather than skill development?

Research and dialogue on the obstacles and advantages of this approach are definitely needed to achieve some balance and to assist professional educators in addressing these issues when using the rubric within their own courses. Another dimension in need of further investigation would be to obtain evidence of convergent and discriminant validity. Finally, a rigorous, larger scale validation study of the universality of the criteria is also warranted if the scale is to become a widespread, valuable and valued tool in the assessment of postsecondary academic skills.

Footnotes

[1] See http://ncme.ed.uiuc.edu/pubs/jem_policy.ace and http://ericae.net/pare/Review.htm

References

Arter, J. (1993). *Designing scoring rubrics for performance assessments: The heart of the matter*. Portland, OR.: Northwest Regional Education Laboratory. (ERIC Document Reproduction Service No. ED 358 143)

Biggs, J. (1995). Assessing for learning: Some dimensions underlying new approaches to educational assessment. *The Alberta Journal of Educational Research, XLI*(1), 1-17.

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E.L., & Lewis Harris, E. (1995/1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress achievement level descriptors as characterizations of mathematics performance. *Educational Assessment, 3*(1), 9-51.

Forgette-Giroux, R., & Simon, M. (1998). L'application du dossier d'apprentissage au niveau universitaire. *Mesure et évaluation en éducation*, 20(3), 85-103.

Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10). Available online: http://ericae.net/pare/getvn.asp?v=7&n=10

Sadler, R. (1989). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191-209.

Simon, M., & Forgette-Giroux, R. (2000). Impact of a content selection framework on portfolio assessment at the classroom level. *Assessment in Education: Principles, Policy and Practice, 17*(1), 103-121.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco: Jossey-Bass Publishers.

APPENDIX

Descriptive scale: EDU5499 Current methods of student assessment in teaching and learning (graduate level course).

Learning objective: To be able to critically analyze the technical qualities of own assessment approaches.

Name: ____________________

Date: ______________

Assessment ☐ formative
☐ summative

Assessor: ☐ self
☐ professor

CRITERIA	PERFORMANCE LEVELS (ANCHORS)			
	GOOD	VERY GOOD	EXCELLENT	EXCEPTIONAL
RELEVANCE Portfolio components address directly the learning objective.	Portfolio components are not necessarily related to the learning objective.	Portfolio components are somehow linked to the learning objective.	Portfolio components are directly linked to the learning objective.	The portfolio reflects exemplary high relevance to the learning objective
SCOPE All aspects of the learning objective and recommended readings are covered within the portfolio.	The portfolio partially reflects the various components of the learning objective.	Most elements of the learning objective and recommended readings are covered within the portfolio.	All aspects of the learning objective and all recommended readings have been dealt with within the portfolio.	The portfolio incorporates treatment of elements beyond the scope of the class and recommended references.
ACCURACY Current concepts, terms, principles, and conventions are used correctly and with clarity throughout the portfolio.	Learned concepts, terms, principles, and conventions are more or less used correctly throughout the portfolio.	The portfolio shows precision in the use of current concepts, terms and principles but relevant conventions not always followed.	The portfolio reflects correct and clear use of terms, concepts, principles, and conventions.	The portfolio demonstrates clear, correct, precise, and concise use of terms, concepts, principles and conventions.

COHERENCE Elements within and across the portfolio are logically and structurally linked together. Ideas are interconnected and are presented in a consistent fashion throughout the portfolio.	Elements and ideas are presented in a disconnected, rather piecemeal fashion.	Elements are somehow linked together but reflect some inconsistency across the portfolio.	Evidence of structural and internal consistency within and to some extent, across the portfolio.	The portfolio is highly and tightly organized. Ideas, concepts and principles are presented in a consistent fashion across the portfolio.
DEPTH The portfolio reflects a personal position supported by a rich analysis of relevant and high quality references.	The portfolio presents a position highly dependent on a superficial analysis of references.	The portfolio presents a position supported by some analysis of relevant references.	The portfolio reflects a personal position based on a deep and thorough analysis of relevant references.	The portfolio presents a personal position based on an integration of relevant and high quality references.

Descriptors: *Rubrics; Scoring; Student Evaluation; Test Construction

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Solomon, David J. (2001). Conducting web-based surveys. *Practical Assessment, Research & Evaluation*, 7(19). Available online: http://ericae.net/pare/getvn.asp?v=7&n=19. This paper has been viewed 1101 times since 8/23/01.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Solomon, David J.

Conducting Web-Based Surveys

David J. Solomon
Office of Medical Education Research and
Development and
the Department of Medicine, College of Human Medicine,
Michigan State University

Web-based surveying is becoming widely used in social science and educational research. The Web offers significant advantages over more traditional survey techniques however there are still serious methodological challenges with using this approach. Currently coverage bias or the fact significant numbers of people do not have access, or choose not to use the Internet is of most concern to researchers. Survey researchers also have much to learn concerning the most effective ways to conduct surveys over the Internet. While in its early stages, research on Internet-based survey methodology has identified a number of factors that influence data quality. Of note, several studies have found Internet surveys have significantly lower response rates than comparable mailed surveys. Several factors have been found to increase response rates including personalized email cover letters, follow-up reminders, pre-notification of the intent to survey and simpler formats. A variety of software tools are now available for conducting Internet surveys and they are becoming a increasing sophisticated and easy to use. While there is a need for caution, the use of Web-based surveying is clearly going to grow.

Introduction

The growth of the Internet has impacted on virtually every aspect of society. Survey research is no exception. Two years ago in an informal search of Yahoo, Kay and Johnson (1999) identified over 2,000 Web-based surveys[1] in 59 areas. The interest in Web-based surveying is not surprising as it offers a number of distinct advantages over more traditional mail and phone techniques. Examples include reducing the time and cost of conducting a survey and avoiding the often error prone and tedious task of data entry (Medin, Roy & Ann, 1999).

Email offers one option for distributing Internet surveys. Up until a few years ago email surveys were the predominate means of Internet surveying. As the World Wide Web (WWW) has grown in popularity, the use of Hypertext Markup Language (HTML) forms or Web-based surveys are becoming the dominant method of gathering survey data. These forms streamline the data collection process formatting and entering responses directly into a database for analysis. Since HTML forms can be made programmable, it is also possible to have real time error checking and correction increasing the accuracy of the data collection process. The formatting capabilities of HTML allow the creation of easy-to-read and attractive forms that may improve response rates. In addition, the programmability of HTML forms makes it possible to randomly order responses and tailor options based on information the respondent supplies earlier in the survey.

Combining an email "cover letter" as a means of contacting sampled people with the use of an HTML form for data collection provides an especially effective and efficient approach to Internet surveying. Modern email packages automatically convert universal resource locators (URLs) or web-addresses in the text of an email into a hyperlinks. Placing the URL of the survey form in a cover letter email allows the respondent to "click" their mouse on the URL to display the survey form and subsequently fill it out.

Concerns with Web-based Surveying

Although Web-based surveying is very attractive, at this point it should be used with caution. Currently the biggest concern in Internet surveying is coverage bias or bias due to sampled people not having or choosing not to access the Internet (Kay & Johnson, 1999; Crawford, Couper & Lamias, 2001). Despite expediential growth of the Internet there are still large numbers of people who do not have access and/or choose not to use the Internet. It is also clear that there are wide disparities in Internet access among ethnic and socioeconomic groups (Selwyn & Robson, 1998).

There are specific populations where Internet access is extremely high and coverage bias is likely to be less of a concern. College students and university faculty within the USA, Canada and Western Europe are examples of such populations. Even though coverage bias may be less of an issue in these groups, experience and comfort with Internet-based tools such as Web browsers is another serious potential source of bias both in response rates and the way people respond to the survey (Dillman, Tortora & Bowker, 2001).

Web-based surveying is still in the early stages of development. The WWW is a unique media and it is not clear to what extent the knowledge we have gained over years of experience with more traditional surveying techniques fully applies to Internet surveying (Dilman, Tortora & Bowker, 2001). Studies are just beginning to done to learn the optimal ways to structure and format Internet surveys to limit biases and increase response rates. It is also likely that the best way to design an Internet survey depends in part on the familiarity and comfort of the respondents in using Web browsers and email clients. It is also quite likely that the type of Internet connection as well as the hardware and software used in accessing the Internet will impact on response rates and possibly how a person responds to an Internet-based survey.

The use of HTML forms for surveying poses a unique set of issues and challenges that need to be addressed to ensure valid data. The Web is a very public place and unless steps are taken to limit access to a survey, it may be found and responded to by people who are not among those sampled by the researcher. This can either happen by accident or maliciously. Since one only has to "click" their mouse pointer on the "submit" button to respond to a Web-based survey instrument once it is filled out, it is also quite possible for respondents to either mistakenly or purposefully submit multiple copies of their responses.

While Internet-based surveying techniques need to be used with caution, their benefits warrant continued exploration and the cautious use. It is also pretty clear that coverage bias and familiarity with Internet tools will be less and less of an issue over time. Additionally our knowledge about how best to conduct Internet surveys will continue to improve with research and experience.

Research on Internet-Based Surveying

Although the research on Internet-based surveying is limited, findings are beginning to appear in the literature. Several studies have found that response rates for Internet surveys are lower that equivalent mail surveys (Medin, Roy & Ann, 1999; Cooper, Blair & Triplett, 1999). As noted by Crawford and colleagues (2001), this may be due to our lack of knowledge on how to achieve high response rates using the Internet surveys. The lower response rates for internet surveys may also reflect coverage bias, the lack of familiarity with the media and/or lack of convenient access to the Internet. In the author's experience, Web congestion can also be a factor in lowering response rates for Web surveys particularly with people who have relatively little experience with the Internet.

Cook and colleagues (2000) conducted a meta-analysis of factors influencing response rates in Internet-based surveys[2]. They found that follow-up contacts with non-respondents, personalized contacts, and contacting sampled people prior to sending out the survey were the three dominant factors in higher response rates.

Kittleson (1997) in a study of email-based surveying found it was possible to double the response rate with follow-up memos though in general this may be somewhat optimistic. As with mailed surveys, repeated follow-ups have diminishing returns and at some point risk irritating potential respondents without noticeably increasing response rates. Additionally, Dillman, Tortora, Conrad & Bowker (2001) found that relatively plain Web surveys that load quickly resulted in higher response rates than "fancier" surveys that take longer to load. Jeavons (1998) analyzed detailed server logs from three separate large-scale surveys. He found a relatively high percentage of potential respondents stopped completing the surveys 1) when encountering the first question, 2) when encountering a complex question grid, and 3) when asked to supply their email address. This suggests that some potential respondents have difficulty with the media and give up early in the process of completing the survey or when encountering complex questions. Others may be reluctant to give out personal information such as an email address. The logs were also merged with demographic data collected via the surveys. Somewhat surprisingly no patterns in failure to complete rates were found by gender, age, or education level. In two of the surveys, people with lower income were found to have a higher rate of repeating screens of questions mainly due to improperly filling out questions.

Developing Web Surveys

As noted, due to their inherent advantages, most Internet surveying is now being done using HTML forms with potential respondents often contacted via email cover letters. While some developers still directly code these forms in HTML, there are dozens of HTML editors available, and they are becoming increasingly sophisticated and easy to use. There are two general methods of capturing the data entered into an HTML form. The form can be programmed to email the data back to a specified email address or captured by a program on the server called a common gateway interface (CGI) script. Using CGI scripts is more robust, offers more flexibility and is the far more commonly used method of capturing data. There are several HTML development packages that both provide HTML editing capabilities and automate the process of developing the CGI scripts necessary to capture data from HTML forms developed with the package. Two widely used examples of these packages are Microsoft's FrontPage and Macromedia's ColdFusion.

While these packages are general purpose Web development tools, there are also a growing number of software development systems designed specifically for Web-based surveying. Examples include Perseus's Survey Solutions for the Web, Creative Research System's The Survey System, and Survey Said™ Survey Software. These packages tend to offer additional features specific to survey research. Examples include as managing the distribution of email cover letters, built-in statistical analysis and reporting capabilities, and automatic tracking of people who have responded coupled with the ability of sending out follow-up email reminders to those who have yet to respond. Their HTML editors are also geared for survey form development, allowing them to simplify and streamline the process of developing and formatting the question response fields.

Web Survey Mailer System

The author has developed a set of software tools that provides many of the complex Web survey administration functions included in Web surveying packages[3]. The software, Web Survey Mailer System (WSMS), is an integrated survey administration system that will send out personalized email cover letters, track which of the sampled people have completed the survey while keeping their responses anonymous and send out subsequent reminder emails to only those sampled people who have not responded to the survey. WSMS will block people who have not been sampled from accessing and responding to the survey and will keep respondents from submitting more than one set of survey responses. The system includes a customizable CGI script to capture the survey responses and place them in a tab-delimited ASCII database format that can easily be downloaded from the server and imported into a standard PC data base or statistical package. WSMS is written in PHP and uses the MySQL relational database to store information on the sampled people. Both PHP and MySQL are stable and powerful "open source" packages widely available on university and commercial Web servers and can be obtained free of charge in a variety of versions that will run on most common server operating systems. The WSMS scripts and documentation are available free-of-charge and can be downloaded from http://www.med-ed-online.org/rsoftware.htm#wsms

Summary

Internet surveys are clearly going to continue to grow in popularity as the problems of coverage bias and unfamiliarity with the Internet subside. For the foreseeable future there will be people who will lack Internet access either by choice or circumstance though this will be less and less of an issue. Additionally the tools for conducting Web-based surveys will continue to grow in sophistication and ease of use as will our knowledge on how best to employ this survey methodology. At present researchers should use this technique with caution in carefully chosen populations and with an eye to learning as much as possible about how to do it better.

Notes

1. In this paper we use the term "Internet survey" for both email and HTML form-based surveying while the term "Web-based survey" is reserved for HTML form-based surveys.

2. Cook, Heath, & Thompson (2000) included studies of both Web- and email-based surveys.

3. Detailed documentation for the Web Survey Mailer System is provided. However, installing and using these tools requires a good working knowledge of HTML and some background and understanding of server-based programming.

References

Best, S.,J., Krueger, B., Hubbard, C. & Smith, A. (2001) An assessment of the generalizability of internet surveys. *Social Science Computer Review*, 19, 131-145.

Cook, C, Heath, F, & Thompson, R. (2000) A meta-analysis of response rates in web or Internet-based surveys. *Educational and Psychological Measurement*, 60, 821-836.

Couper, M.P., Blair, J. & Triplett T (1999) A comparison of mail and e-mail for a survey of employees in federal statistical agencies. *Journal of Official Statistics,* 15, 39-56.

Crawford, S.D., Couper, M.P.& Lamias, M.J.. (2001) Web Surveys: Perception of burden. *Social Science Computer Review*, 19, 146-162.

Dillman, D.A., Tortora, R.D. & Bowker, D. Principles for constructing web surveys. Working paper available from http://survey.sesrc.wsu.edu/dillman/papers.htm [Accessed 6/01]

Dillman, D.A, Tortora, R.D, Conrad, J. & Bowker D. Influence of plan vs. fancy design on response rates of Web surveys. Working paper available from http://survey.sesrc.wsu.edu/dillman/papers.htm [Accessed 6/01]

Jeavons A. Ethology and the Web: Observing respondent behavior in Web surveys. Proceedings of the Worldwide Internet Conference, Amsterdam: ESOMAR, 1998, available from http://w3.one.net/~andrewje/ethology.html [Accessed 7/01]

Kaye B.K. & Johnson T.J. (1999) .Research Methodology: Taming the Cyber Frontier. *Social Science Computer Review*, 17, 323-337.

Kittleson, M. (1997) Determining effective follow-up of e-mail surveys. *American Journal of Health Behavior*. 21, 193-196.

Medin, C., Roy, S. & Ann, T. (1999) World Wide Web versus mail surveys: A comparison and report.. Paper presentation at ANZMAC99 Conference, Marketing in the Third Millennium, Sydney, Australia, available from http://www.anzmac99.unsw.edu.au/anzmacfiles/papers.htm [Accessed 6/01]

Selwyn, N., Robson, K. (1998) Using e-mail as a research tool, *Social Research Update*, available from http://www.soc.surrey.ac.uk/sru/SRU21.html [Accessed 6/01]

Descriptors: *World Wide Web; *Survey Methods; Response Rates [Questionnaires]; *Surveys; Electronic Mail

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Cassady, Jerrell C. (2001). The stability of undergraduate students' cognitive test anxiety levels. *Practical Assessment, Research & Evaluation*, 7(20). Available online: http://ericae.net/pare/getvn.asp?v=7&n=20. This paper has been viewed 707 times since 8/23/01.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

Find articles in ERIC written by
Cassady, Jerrell C.

The Stability of Undergraduate Students' Cognitive Test Anxiety Levels

Jerrell C. Cassady
Department of Educational Psychology
Ball State University

Test anxiety has been overwhelmingly identified as a two-factor construct, consisting of the cognitive (often referred to as "worry") and emotional (or affective) components (Morris, Davis, & Hutchings, 1981; Schwarzer, 1986). The predominant view of the relationship between these two factors suggests the cognitive component directly impacts performance (Bandalos, Yates, & Thorndike-Christ, 1995; Cassady & Johnson, in press; Hembree, 1988), while the emotionality component is related but does not directly influence test performance (Sarason, 1986; Williams, 1991). The apparent relationship between emotionality and test performance is such that emotionality impacts test performance only under situations where the individual also maintains a high level of cognitive test anxiety (Deffenbacher, 1980; Hodapp, Glanzmann, & Laux, 1995). Although emotionality has traditionally not been viewed as central to performance, recent work has demonstrated that emotionality may be the triggering mechanism for self-regulation strategies that facilitate performance (Schutz & Davis, 2000).

This study investigated the stability of test anxiety over time by examining the level of reported cognitive test anxiety at three points in an academic semester (all proximally close to exams). The comparisons between levels of test anxiety over time were intended to track fluctuations in level of anxiety over time and across testing formats. The expectation was that cognitive test anxiety is a relatively stable (trait-like) construct, and that students' levels of anxiety would reflect a high degree of similarity within subjects over time. The underlying purpose of the study was to determine if it is indeed necessary to evaluate levels of test anxiety for each test taken, or if test anxiety is stable enough that evaluation at one point in time is sufficient for research that spans multiple exams. Because stability is largely influenced by the internal consistency or reliability of the measure in, it was also imperative to investigate the level of scale reliability at each point in the study. Under conditions where a scale is (a) internally consistent (reliable), and (b) demonstrates high levels of similarity in response over extended periods of time, it is reasonable to establish that the measured construct is stable (Nunnally & Bernstein, 1994).

Standard conceptualizations of the cognitive test anxiety construct have addressed the interplay between state and trait anxiety (Snow, Corno, & Jackson, 1996; Spielberger & Vagg, 1992). In this conceptualization, individuals with high levels of cognitive test anxiety generally hold heightened levels of trait anxiety, but in evaluative situations, their state anxiety also elevates (Zeidner, 1995). This combinatory relationship can lead to feelings of anxiety that interfere with test performance through blocks to cue utilization, attenuated attentional resources, or mere cognitive interference from the worries and fears induced by anxiety (Geen, 1980; Hembree, 1988; Sarason, 1986). This relationship has also been characterized as an additive function of the dispositional and situational anxiety influences faced by students in evaluative scenarios (Zohar, 1998).

The attention to the situation-specific factors that lead to test anxious thoughts and behaviors has promoted the methodological practice of gathering test anxiety data as close to the testing event as possible, to best capture the contextual influences to anxiety in the research design (Cassady & Johnson, in press; Covington, 1985; Hodapp et al., 1995; Zeidner, 1998). Thus, the ideal time to test for test anxiety would be during an examination itself, with the subject providing "online" responses to the immediate feelings, fears, and behavioral responses that are arising during evaluation. However, this cannot occur in research designs that use students in actual testing situations, due to the probability of inducing additional debilitative cognitive test anxiety by having the student respond to items addressing their level of worry or fear for tests. Further, the implication is that in order to measure the performance-test anxiety relationship over a series of tests would require repeated administration of the test anxiety measures. These conditions are pragmatically undesirable, and potentially unnecessary.

Method

Participants

Undergraduate students in an introductory educational psychology course were the participants in this investigation. The participants completed the study instruments during one academic semester, as one option for completing course credit. Sixty-four undergraduate students participated in the three phases of study, with several participants completing only portions of the data. The participants were predominantly White (n = 62), with two Black students participating (which included all Black students available for participation). Consistent with the course population, there were 47 females and 17 males.

Materials

All measures in this study have been validated and found to have high levels of internal consistency (Cassady, 2001b; Cassady & Johnson, in press). However, there has been no investigation with the Cognitive Test Anxiety scale to identify the test-retest reliability, or the stability of test anxiety over time with this instrument.

The Cognitive Test Anxiety scale (Cassady & Johnson, in press) is a 27-item measure focused on only the cognitive domain of test anxiety, formerly referred to as worry. The cognitive domain includes the tendency to engage in task-irrelevant thinking during test taking and preparation periods, the tendency to draw comparisons to others during test taking and preparation periods, and the likelihood to have either intruding thoughts during exams and study sessions, or have relevant cues escape the learner's attention during testing. In previous investigations of the Cognitive Test Anxiety scale, involving over 1,000 participants, a reliable method for determining high and low levels of test anxiety has been documented (Cassady, 2001a; 2001b; Cassady & Johnson, in press), which splits respondents into three levels of test anxiety. The low group scores range from 27 to 61, moderate test anxiety group ranges from 62 to 71, and responses that score over 72 are categorized as the high test anxiety group (maximum score possible = 108).

Sarason's (1984) Bodily Symptoms subscale from the Reactions to Tests four-factor scale of test anxiety was used as a measure of the emotionality component of test anxiety. The Bodily Symptoms scale has been shown to have an acceptable degree of internal consistency, despite the short length (k = 10; Sarason, 1984; 1986).

Procedures

Throughout the course of one academic semester, students were invited to complete the Cognitive Test Anxiety scale and the Bodily Symptoms subscale three times. The timing of completion of these scales was aligned with the testing times in the course. All students completed the scales no more than seven days before they took the course examination. The time was variable due to the students' ability to choose which day to complete the examination. All scales were completed in groups, in the students' course classroom. Scale completion generally took between 8 and 15 minutes.

The first two course exams were 35-item multiple-choice tests, administered online at the students' convenience. However, the third exam was a take-home, open book exam that had the same weight in overall course grade as the other two exams. Due to the methods of data collection that were designed to maintain student confidentiality as related to the test anxiety and emotionality responses, matching students test scores to the dependent variables in this study was not possible.

Results

Data analyses focused on examining the stability of the students' cognitive test anxiety reports, and their perceptions of emotionality as measured by bodily symptoms. Correlation analyses were used to illustrate the students' stable reactions to the test events posed in the course. In addition, internal consistency for each scale was calculated to identify the level of reliability demonstrated in each administration of the two dependent measures. Finally, correction for attenuation of the correlations among the measures was conducted to provide an analysis of the hypothetical true score correlations given the condition that no measurement error were present in the study measures (Nunnally & Bernstein, 1994).

Initially, descriptive analyses on the test anxiety scores demonstrated that the students in this course had somewhat lower levels of test anxiety than previous uses of the scale (Cassady & Johnson, in press; Cassady 2001a; 2001b). The average cognitive test anxiety scores were similar for test 1 (M = 62.44, SD = 14.41, n = 59), test 2 (M = 62.46, SD = 15.39, n = 57), and test 3 (M = 61.49, SD = 14.94, n = 57), which all place the average score at the pre-established cut-off point between low and moderate cognitive test anxiety (Cassady, 2001a; 2001b). The mean scores on the bodily symptoms measure also were stable across the three administrations, and were somewhat lower than average. With a possible score range of 10 to 40, the average scores were in the low range for test 1 (M = 15.73, SD = 5.10, n = 56), test 2 (M = 16.41, SD = 5.73, n = 56), and test 3 (M = 15.48, SD = 5.62, n = 56).

To measure the stability of cognitive test anxiety and emotionality over the course of a semester, correlational analyses of the three points of administration were conducted (see Table 1). The results demonstrate that there is a very strong correlation between the students' reports of cognitive test anxiety across three points in the semester, as well as strong correlations among the three emotionality measurements. Further, the correlations between test anxiety and bodily symptoms are significant, and consistent with earlier research on the relationship between the two primary factors of test anxiety (Hembree, 1988). The correlations between cognitive test anxiety and bodily symptoms did reveal that the bodily symptoms scale administration that took place before the second exam was the strongest correlation for all three administrations of the Cognitive Test Anxiety scale. This deviates from the expectation that suggests contextual factors in place at each administration session would drive the cognitive and emotional components given together to be most similar. Although these correlational values do not vary greatly, the pattern may be due to the fact that the students adjusted their reports of emotionality in response to the first course examination. This is supported by the fact that the second test administration period was also the point at which the Bodily Symptoms subscale mean was highest for the population.

Table 1
Cognitive Test Anxiety and Bodily Symptoms Intercorrelation Matrix

Measure	1	2	3	4	5	6
1. CTA (Test 1)[1]	.94 (59)	.96	.94	.63	.70	.51
2. CTA (Test 2)	.91 (52)	.95 (57)	.98	.56	.73	.54
3. CTA (Test 3)	.88 (53)	.93 (55)	.94 (57)	.52	.70	.58
4. BS (Test 1)[2]	.58 (56)	.52 (50)	.48 (50)	.91 (56)	.92	.87
5. BS (Test 2)	.64 (51)	.67 (56)	.64 (54)	.82 (49)	.88 (56)	.94
6. BS (Test 3)	.47 (52)	.50 (54)	.54 (56)	.79 (49)	.84 (53)	.91 (56)

Note: All p's < .001. Values on the diagonal indicate Cohen's Alpha Inter-item reliability coefficient. Values on the top half of the table reflect the correlations after correction for attenuation, while the bottom half are not corrected. Values in parentheses report sample size for each analysis.

[1] Cognitive Test Anxiety scale score.

[2] Bodily Symptoms subscale score.

Finally, the internal consistencies of the two dependent measures were reported (see Table 1) for each of the three administration periods, using Cronbach's Alpha (Nunnally & Bernstein, 1994). These measures of internal consistency confirmed previous work that demonstrated high levels of internal reliability, and were subsequently used to correct the correlational values for attenuation due to measurement error. The corrections for attenuation did not reveal any changes to the patterns of intercorrelations.

Discussion

The results demonstrate that it is methodologically practical to make use of test anxiety data gathered at times other than when a particular test in question is being completed, given that the data have been collected during a time in which typical test-induced contextual variables are activated. That is, it does not appear to be important to gather test anxiety data prior to every test upon which the anxiety construct is being used in research analyses, simply prior to any test. The high internal consistency paired with the high degree of correlation among repeated administrations of the Cognitive Test Anxiety scale and Bodily Symptoms subscale provide strong evidence that the constructs have long-range stability (Nunnally &

Bernstein, 1994).

The data provide useful information regarding an efficient and methodologically sound approach for collecting test anxiety data from undergraduate students. It is reasonable to extrapolate from these results that test anxiety data collected in close proximity to an evaluative event can be used in analyses of the impact of test anxiety on any test within that academic period. The data also demonstrated that the Cognitive Test Anxiety scale, which has been shown to have high levels of internal consistency and high construct validity (Cassady & Johnson, in press), also provides stable and consistent measures of test anxiety over time and across testing formats.

One theoretical implication of these results also relates to the interpretation of test anxiety as a failure at multiple levels of information processing (Benjamin, McKeachie, Lin, & Holinger, 1981; Naveh-Benjamin, 1991; McKeachie, 1984). The research in this area has confirmed that students with high test anxiety are not only prone to failure in situations where time factors attenuate performance and retrieval of key information, but even in take-home examinations (Benjamin et al., 1981). Although exam performances were not available for this study, and no conclusions regarding the stability of a detrimental impact of test anxiety on performance can be drawn, the results demonstrate that the level of reported anxiety is consistent over time, despite variations in course exam format. That is, the level of anxiety induced by the take-home examination did not differ significantly from the level of anxiety induced by the closed book multiple-choice examinations. Therefore, it seems that test anxious thoughts and behaviors are likely prompted by the presence of evaluative tasks, regardless of testing format. This finding extends previous work demonstrating no differential rates of cognitive test anxiety induced by in-class and online testing formats (Cassady, 2001a).

References

Bandalos, D. L., Yates, K., & Thorndike-Christ, T. (1995). Effects of math self-concept, perceived self-efficacy, and attributions for failure and success on test anxiety. *Journal of Educational Psychology, 87*, 611-623.

Benjamin, M., McKeachie, W. J., Lin, Y., & Holinger, D. P. (1981). Test anxiety: Deficits in information processing. *Journal of Educational Psychology, 73*, 816-824.

Cassady, J. C. (2001a). The effects of online formative and summative assessment on undergraduate students' achievement and cognitive test anxiety. Manuscript submitted for publication.

Cassady, J. C. (2001b). Cognitive test anxiety in undergraduate students in Kuwait and the United States. Manuscript submitted for publication.

Cassady, J. C., & Johnson, R. E. (in press). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology.*

Covington, M. V. (1985). Test anxiety: Causes and effects over time. In H. M. van der Ploeg, R. Schwarzer, & C. D. Spielberger (Eds.) *Advances in Test Anxiety Research (Vol. 4)* (pp. 55-68). Lisse, The Netherlands: Swets & Zeitlinger.

Deffenbacher, J. L. (1980). Worry and emotionality in test anxiety. In I. G. Sarason, (Ed.) *Test anxiety: Theory, research, and applications* (pp. 111-124). Hillsdale, NJ: Lawrence Erlbaum.

Geen, R. G. (1980). Test anxiety and cue utilization. In I. G. Sarason (Ed.), *Test anxiety: Theory, research, and applications* (pp. 43-61). Hillsdale, NJ: Lawrence Erlbaum.

Hembree, R. (1988). Correlates, causes, and treatment of test anxiety. *Review of Educational Research, 58,* 47-77.

Hodapp, V., Glanzmann, P. G., & Laux, L. (1995). Theory and measurement of test anxiety as a situation-specific trait. In C. D. Spielberger & P. R. Vagg (Eds.) *Test anxiety: Theory, assessment, and treatment* (pp. 47-59). Washington, D.C.: Taylor & Francis.

McKeachie, W. J. (1984). Does anxiety disrupt information processing or does poor information processing lead to anxiety? *International Review of Applied Psychology, 33,* 187-203.

Morris, L. W., Davis, M. A., & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale. *Journal of Educational Psychology, 73,* 541-555.

Naveh-Benjamin, M. (1991). A comparison of training programs intended for different types of test-anxious students: Further support for an information-processing model. *Journal of Educational Psychology, 83,* 134-139.

Naveh-Benjamin, M., McKeachie, W. J., & Lin, Y. (1987). Two types of test-anxious students: Support for an information processing model. *Journal of Educational Psychology, 79,* 131-136.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3^{rd} Ed.).* New York: McGraw-Hill.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to Tests. *Journal of Personality and Social Psychology, 46,* 929-938.

Sarason, I. G. (1986). Test anxiety, worry, and cognitive interference. In R. Schwarzer (Ed.) *Self-related cognitions in anxiety and motivation* (pp. 19-34). Hillsdale, NJ: LEA.

Schutz, P. A., & Davis, H. A. (2000). Emotions and self-regulation during test taking. *Educational Psychologist, 35,* 243-256.

Schwarzer, R. (1986). Self-related cognitions in anxiety and motivation: An introduction. In R. Schwarzer (Ed.), *Self-related cognitions in anxiety and motivation* (pp. 1- 18). Hillsdale, NJ: LEA.

Snow, R. E., Corno, L., & Jackson, D. (1996). Individual differences in affective and conative functions. In D. C. Berliner & R. C. Calfee (Eds.) *Handbook of educational psychology* (pp. 243-310). New York: Macmillan.

Spielberger, C. D., & Vagg, P. R. (1995). Test anxiety: A transactional process model. In C. D. Spielberger & P. R. Vagg (Eds.) *Test anxiety: Theory, assessment, and treatment* (pp. 1-14). Washington, D.C.: Taylor & Francis.

Williams, J. E. (1991). Modeling test anxiety, self concept and high school students' academic achievement. *Journal of Research and Development in Education, 25,* 51-57.

Zeidner, M. (1995). Adaptive coping with test situations: A review of the literature. *Educational Psychologist, 30(3),* 123-133.

Zeidner, M. (1998). *Test anxiety: The state of the art.* New York: Plenum Press.

Zohar, D. (1998). An additive model of test anxiety: Role of exam-specific expectations. *Journal of Educational Psychology, 90,* 330-340.

Descriptors: *Test Anxiety; *Test Reliability; Test Construction; Test Validity

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

La Marca, Paul M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, **7**(21). Available online: http://ericae.net/pare/getvn.asp?v=7&n=21. This paper has been viewed 1039 times since 9/17/01.

‣ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

‣ Find articles in ERIC written by
La Marca, Paul M.

Alignment of Standards And Assessments as an Accountability Criterion

Paul M. La Marca
Nevada Department of Education

To make defensible accountability decisions based in part on student and school-level academic achievement, states must employ assessments that are aligned to their academic standards. Federal legislation and Title I regulations recognize the importance of alignment, which constitutes just one of several criteria for sound assessment and accountability systems. However, this seemingly simplistic requirement grows increasingly complex as its role in the test validation process is examined.

This paper provides an overview of the concept of alignment and the role it plays in assessment and accountability systems. Some discussion of methodological issues affecting the study of alignment is offered. The relationship between alignment and test score interpretation is also explored.

The Concept of Alignment

Alignment refers to the degree of match between test content and the subject area content identified through state academic standards. Given the breadth and depth of typical state standards, it is highly unlikely that a single test can achieve a desirable degree of match. This fact provides part of the rationale for using multiple accountability measures and also points to the need to study the degree of match or alignment both at the test level and at the system level. Although some degree of match should be provided by each individual test, complementary multiple measures can provide the necessary degree of coverage for systems alignment. This is the greater accountability issue.

Based on a review of literature (La Marca, Redfield, & Winter 2000), several dimensions of alignment have been identified. The two overarching dimensions are content match and depth match. Content match can be further refined into an analysis of broad content coverage, range of coverage, and balance of coverage. Both content and depth match are predicated on item-level comparisons to standards.

Dimensions of Alignment

Content Match. *How well does test content match subject area content identified through state academic standards?*

- Broad content coverage. *Does test content address the broad academic standards? Is there categorical congruence?*
- Range of coverage. *Do test items address the specific objectives related to each standard?*
- Balance of coverage. *Do test items reflect the major emphases and priorities of the academic standards?*

Depth Match. *How well do test items match the knowledge and skills specified in the state standards in terms of cognitive complexity? A test that emphasized simple recall, for example, would not be well-aligned with a standard calling for students to be able to demonstrate a skill.*

Broad content match, labeled categorical congruence by Webb (1997), refers to alignment at the broad standard level. For example, a general writing standard may indicate that "students write a variety of texts that inform, persuade, describe, evaluate, or tell a story and are appropriate to purpose and audience " (Nevada Department of Education, 2001 p. 14). Obviously this standard covers a lot of ground and many specific indicators of progress or objectives contribute to attainment of this broadly defined skill. However, item/task match at the broad standard level can drive the determination of categorical congruence with little consideration to the specific objectives being measured.

As suggested above, the breadth of most content standards is further refined by the specification of indicators or objectives. Range of coverage refers to how well items match the more detailed objectives. For example, the Nevada writing standard noted above includes a variety of specific indicators: information, narration, literary analysis, summary, and persuasion. Range of coverage would require measurement to be spread across the indicators. Similarly, the balance of coverage at the objective level should be judged based on a match between emphasis in test content and emphasis prescribed in standards documents.

Depth alignment refers to the match between the cognitive complexity of the knowledge/skill prescribed by the standards and the cognitive complexity required by the assessment item/task (Webb 1997, 1999). Building on the writing example, although indirect measures of writing, such as editing tasks, may provide some subject-area content coverage, the writing standard appears to prescribe a level of cognitive complexity that requires a direct assessment of writing to provide adequate depth alignment.

Alignment can best be achieved through sound standards and assessment development activities. As standards are developed, the issue of how achievement will be measured should be a constant consideration. Certainly the development of assessments designed to measure expectations should be driven by academic standards through development of test blueprints and item specifications. Items/tasks can then be designed to measure specific objectives. After assessments are developed, a post hoc review of alignment should be conducted. This step is important where standards-based custom assessments are used and absolutely essential when states choose to use assessment products not specifically designed to measure their state standards. Whenever assessments are modified or passing scores are changed, another alignment review should be undertaken.

Methodological Consideration

An objective analysis of alignment as tests are adopted, built, or revised ought to be conducted on an ongoing basis. As will be argued later, this is a critical step in establishing evidence of the validity of test score or performance interpretation.

Although a variety of methodologies are available (Webb, 1999; Schmidt, 1999), the analysis of alignment requires a two-step process:

- a systematic review of standards and
- a systematic review of test items/tasks.

This two-step process is critical when considering the judgment of depth alignment.

Individuals with expertise in both subject area content and assessment should conduct the review of standards and assessments. Reviewers should provide an independent or unbiased analysis; therefore, they should probably not have been heavily involved in the development of either the standards or the assessment items.

The review of standards and assessment items/tasks can occur using an iterative process, but Webb (1997, 1999) suggests that the review of standards precede any item/task review. An analysis of the degree of cognitive complexity prescribed by the standards is a critical step in this process. The subsequent review of test items/tasks will involve two decision points

- a determination of what objective, if any, an item measures and
- the items degree of cognitive complexity.

The subjective nature of this type of

Alignment Process

review requires a strong training component. For example, the concept of depth or cognitive complexity will likely vary from one reviewer to the next. In order to code consistently, reviewers will need to develop a shared definition of cognitive complexity. To assist in this process, Webb (1999) has built a rubric that defines the range of cognitive complexity, from simple recall to extended thinking. Making rubric training the first step in the formal evaluation process can help to reinforce the shared definition and ground the subsequent review of test items/tasks.

Conduct a systematic review of standards.

Conduct a systemic review of test items/tasks:

- Determine what objective(s) each item/task measures.
- Determine the degree of each item's cognitive complexity.

Systematic review of standards and items can yield judgments related to broad standard coverage, range of coverage, balance of coverage, and depth coverage. The specific decision rules employed for each alignment dimension are not hard and fast. Webb (1999) does provide a set of decision rules for judging alignment and further suggests that determination of alignment should be supported by evidence of score reliability.

Thus far the discussion has focused on the evaluation of alignment for a single test instrument. If the purpose of the exercise is ultimately to demonstrate systems alignment, the process can be repeated for each assessment instrument sequentially, or all assessment items/tasks can be reviewed simultaneously. The choice may be somewhat arbitrary. However, there are advantages to judging alignment at both the instrument level and the system level. If, for example, decisions or interpretations are made based on a single test score, knowing the test's degree of alignment is critical. Moreover, as is typical of school accountability models, if multiple measures are combined prior to the decision-making or interpretive process, knowledge of overall systems alignment will be critical.

Why is alignment a key issue

In the current age of educational reform in which large-scale testing plays a prominent role, high-stakes decisions predicated on test performance are becoming increasingly common. As the decisions associated with test performance carry significant consequences (e.g., rewards and sanctions), the degree of confidence in, and the defensibility of, test score interpretations must be commensurably great. Stated differently, as large-scale assessment becomes more visible to the public, the roles of reliability and validity come to the fore.

Messick (1989) has convincingly argued that validity is not a quality of a test but concerns the inferences drawn from test scores or performance. This break from traditional conceptions of validity changes the focus from establishing different sorts of validity (e.g., content validity vs. construct validity) to establishing several lines of validity evidence, all contributing to the validation of test score inferences.

Alignment as discussed here is related to traditional conceptions of content validity. Messick (1989) states that "Content validity is based on professional judgments about the relevance of the test content to the content of a particular behavioral domain of interest and about the representativeness with which item or task content covers that domain" (p. 17). Arguably, the establishment of evidence of test relevance and representativeness of the target domain is a critical first step in validating test score interpretations. For example, if a test is designed to measure math achievement and a test score is judged relative to a set proficiency standard (i.e., a cut score), the interpretation of math proficiency will be heavily dependent on a match between test content and content area expectations.

Moreover, the establishment of evidence of content representativeness or alignment is intricately tied to evidence of construct validity. Although constructs are typically considered latent causal variables, their validation is often captured in measures of internal and external structure (Messick, 1989). Arguably the interpretation of measures of internal consistency and/or factor structures, as well as associations with external criterion, will be informed by an analysis of range of content and balance of content coverage.

Therefore, alignment is a key issue in as much as it provides one avenue for establishing evidence for score interpretation. Validity is not a static quality, it is "an evolving property and validation is a continuing process" (Messick, p. 13). As argued earlier, evaluating alignment, like analyzing internal consistency, should occur regularly, taking its place in the cyclical process of assessment development and revision.

Discussion

Alignment should play a prominent role in effective accountability systems. It is not only a methodological requirement but also an ethical requirement. It would be a disservice to students and schools to judge achievement of academic expectations based on a poorly aligned system of assessment. Although it is easy to agree that we would not interpret a student's level of proficiency in social studies based on a math test score, interpreting math proficiency based on a math test score requires establishing through objective methods that the math test score is based on performance relative to skills that adequately represent our expectations for mathematical achievement.

There are several factors in addition to the subjective nature of expert judgments that can affect the objective evaluation of alignment. For example, test items/tasks often provide measurement of multiple content standards/objectives, and this may introduce error into expert judgments. Moreover, state standards differ markedly from one another in terms of specificity of academic expectations. Standards that reflect only general expectations tend to include limited information for defining the breadth of content and determining cognitive demand. Not only does this limit the ability to develop clearly aligned assessments, it is a barrier to the alignment review process. Standards that contain excessive detail also impede the development of assessments, making an acceptable degree of alignment difficult to achieve. In this case, prioritization or clear articulation of content emphasis will ease the burden of developing aligned assessments and accurately measuring the degree of alignment.

The systematic study of alignment on an ongoing basis is time-consuming and can be costly. Ultimately, however, the validity of test score interpretations depends in part on this sort of evidence. The benefits of confidence, fairness, and defensibility to students and schools outweigh the costs. The study of alignment is also empowering in as much as it provides critical information to be used in revising or refining assessments and academic standards.

References

La Marca, P. M., Redfield, D., & Winter, P.C. (2000). *State Standards and State Assessment Systems: A Guide to Alignment*. Washington, DC: Council of Chief State School Officers.

Messick, S. (1989). Validity. In R. L. Linn (Editor), *Educational Measurement (3rd Edition)*. New York: American Council on Education – Macmillan Publishing Company.

Nevada Department of Education (2001). *Nevada English Language Arts: Content Standards for Kindergarten and Grades 1, 2, 3, 4, 5, 6, 7, 8 and 12.*

Schmidt, W. (1999). Presentation in R. Blank (Moderator), The Alignment of Standards and Assessments. Annual National Conference on Large-Scale Assessment, Snowbird, UT.

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of Science and Mathematics Standards and Assessments in Four States*. Washington, DC: Council of Chief State School Officers.

The author would like to acknowledge Phoebe Winter, Council of Chief State School Officers, and Doris Redfield, Appalachia Educational Laboratory, for their assistance in critiquing this manuscript. I would like to acknowledge the CCSSO SCASS-CAS alignment work group for preliminary work in this area.

Correspondence concerning this article should be addressed to Paul M. La Marca, Director of Standards, Curricula, and Assessments, Nevada Department of Education, 700 E. Fifth St., Carson City, Nevada 89436. Electronic mail may be sent to plamarca@nsn.k12.nv.us.

Descriptors: Academic Standards; Educational Change; Evaluation Methods; Instructional Materials; *Item Analysis; *Accountability; Achievement Gains

Home | Articles | Subscribe | Review | Policies

Volume: 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Taub, Gordon E. (2001). A confirmatory analysis of the wechsler adult intelligence scale-third edition: is the verbal/performance discrepancy justified?. *Practical Assessment, Research & Evaluation*, 7(22). Available online: http://ericae.net/pare/getvn.asp?v=7&n=22. This paper has been viewed 614 times since 9/19/01.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
  ERIC/AE Abstracts In Progress
- Find articles in ERIC written by
  Taub, Gordon E.

A Confirmatory Analysis of the Wechsler Adult Intelligence Scale-Third Edition: Is the Verbal/Performance Discrepancy Justified?

Gordon E. Taub
University of Central Florida

Now, in its third revision, the Wechsler Adult Intelligence Scale-III has a new factor structure. The theoretical structure of the new instrument is hierarchical in nature and contains four first-order factors with a second-order g factor at the apex. In addition to the theoretical hierarchical factor structure of the new instrument, there is an implied factor structure that is used for scoring. This implied structure contains a Full Scale IQ Index, Verbal IQ Index, and a Performance IQ Index. This study investigated the construct validity of both the implied and explicit theoretical structure of the instrument. The results indicate the WAIS-III provides an excellent measure of the four factor model and a general factor. The data, however, did not support the construct validity of a Verbal IQ Index/Performance IQ Index dichotomy. These findings and practical implications for the clinician using the instrument are discussed.

In 1997 The Psychological Corporation introduced the latest version of their adult intelligence test, The Wechsler Adult Intelligence Scale-III (WAIS-III). This instrument was first introduced in 1939 and is currently the most widely used test of adult intelligence. The basic format of the new instrument is very similar to the Wechsler Adult Intelligence Scale-Revised and its predecessor the Wechsler Adult Intelligence Scale. The factor structure of WAIS-III is also similar to earlier editions. One important aspect of the new instrument is the existence of an explicit and implicit first-order factor structure. The instrument's explicit factor structure is hierarchical in nature and contains a second-order general factor at the apex and broad first-order factors. The four broad first-order factors are identified as measures of verbal comprehension, perceptual organization, working memory, and processing speed. The inclusion additional broad first-order factors is more consistent with contemporary theories of intelligence (see Carroll, 1993).

Scores for the implied and theoretical models are calculated by combining scaled scores from the various subtests described in Table 1. The four first-order theoretical factors and their associated subtests are presented in Table 2. The calculation of scores for the four factor model is similar to the method used to calculate Verbal IQ (VIQ) and Performance IQ (PIQ) Index scores on earlier editions. It is still possible, however, to calculate VIQ and PIQ scores on the WAIS-III. Although at first this may make the clinician comfortable with the new instrument, it causes confusion when interpreting the results at a first-order factor level. This is because the publisher does not offer a theoretical justification for a using a dichotomous VIQ/PIQ factor structure. Yet, one must sum a participant's VIQ and PIQ scores to obtain a Full Scale IQ Index score (FSIQ). The use of first-order Verbal and Performance factors and a second-order general factor is considered an implicit factor structure because, as previously discussed, theoretical justification for the use of these factors is not contained in the instrument's technical manual. It appears the publisher may have retained the VIQ/PIQ dichotomy in difference to the Wechsler tradition, in contrast to psychometric theory. Therefore the authors offer the clinician both a theoretical factor structure containing four first-order factors and an implicit model with two first-order factors.

Table 1: Descriptions of the WAIS-III Subtests

Subtest	Description
Information	Samples an individual's fund of knowledge acquired through school and cultural experience.
Vocabulary	Measures an individual's vocabulary.
Similarities	The ability to find and synthesize verbal relationships.
Comprehension	Samples practical information and social knowledge.
Arithmetic	Measures mental concentration and computational skill.
Digit Span	Samples short-term memory by requiring the individual to recall increasingly longer strings of numbers.

Letter-Number Sequencing	Samples sequential processing by requiring an individual to correctly order letters and numbers presented orally.
Picture Arrangement	Samples nonverbal reasoning and planning by arranging pictures to tell a story.
Picture Completion	Samples an individual's attention to detail and visual recognition of objects.
Matrix Reasoning	Samples nonverbal perceptual reasoning by requiring an individual to complete the missing portion of abstract patterns.
Block Design	Samples visual-spatial integration by requiring an individual to reproduce abstract patterns.
Coding	Samples the accuracy and speed of visual motor coordination and scanning ability.
Symbol Search	Measures speed and accuracy and attention.
Object Assembly	Puzzles that form a meaningful whole (optional test)

Table 2: Theoretical Factor Structure of the WAIS-III

Factor	Subtests
Verbal Comprehension	Vocabulary, Similarities, Information, Comprehension
Perceptual Organization	Block Design, Matrix Reasoning, Picture Completion, Picture Arrangement
Working Memory	Digit Span, Arithmetic, Letter-Number Sequencing
Processing Speed	Digit Symbol-Coding, Symbol Search

The first purpose of this study is to investigate the construct validity of the two models. The second purpose is to provide the clinician with guidance to interpret results based on the implicit theoretical model of the WAIS-III and the explicit four factor theoretical model offered by the publisher in the technical manual.

Method

Participants

The sample for this study consisted of the standardization sample of the WAIS-III. The WAIS-III was standardized on 2,450 individuals. Thirteen age levels are represented, ranging from 16 to 89 years (For a description of the entire standardization sample see the WAIS-III Technical Manual, 1997.) The psychometric properties of the WAIS-III subtests have been termed "excellent" (Sattler & Ryan, 1999).

Analyses

The model was estimated using the averaged covariance matrix from ages 16-89 from the standardized data and the sample size was set at 200 for the analysis (the average sample size for each age level; see Keith, 1990; Keith & Witta 1997; and Bickley et al., 1995). Confirmatory factor analysis (CFA) via the AMOS program was used to test the fit of the data to each model (Arbuckle, 1997).

The *explicit* theoretical model is identified in the technical manual and provides the theoretical structure of the WAIS-III's four first-order factors. The *implicit* model specifies the WAIS-III model as portrayed on the test record to calculate FSIQ and therefore includes the appropriate placement of the first-order Verbal/Performance constructs. The Object Assembly subtest was omitted because it is an optional test.

Results and Discussion

In CFA, the factor structure is restricted a priori according to guidelines offered by theory. The obtained data is then compared with the restricted, theoretical model. Chi-square statistics indicate the degree of correspondence, or the "goodness of fit", between a proposed model and the empirical data. A number of indices of fit are reported, as suggested by several researchers (e.g., Keith, 1997). The Tucker-Lewis Index (TLI, also called the non-normed fit index), the Comparison Fit Index (CFI), and the Adjusted Goodness of Fit index (AGFI) provided additional measures of fit. For each of these additional indices of fit, values range from 0 to 1.0, with 0 indicating a poor fit, and 1.0 indicating a perfect fit. Generally, values over .90 are considered excellent. To make comparisons between factor models, chi-squares were compared, with significant reductions in the chi-square indicating a better fit of the data the theoretical model.

Figure 1 displays the factor loadings of the four factor model. Figure 2 contains the factor structure of the implied two factor model. Interestingly, the performance factor has a loading of 1.00 on the second-order factor. This indicates that the variance associated with this factor is completely subsumed by the general factor and eliminating the performance factor from the model may provide an improvement in fit. To test this hypothesis a third analysis was conducted with one intermediate first-order Verbal factor and a second-order general factor. The only difference between the model displayed in Figure 3 and the previous model is the

Performance factor was eliminated.

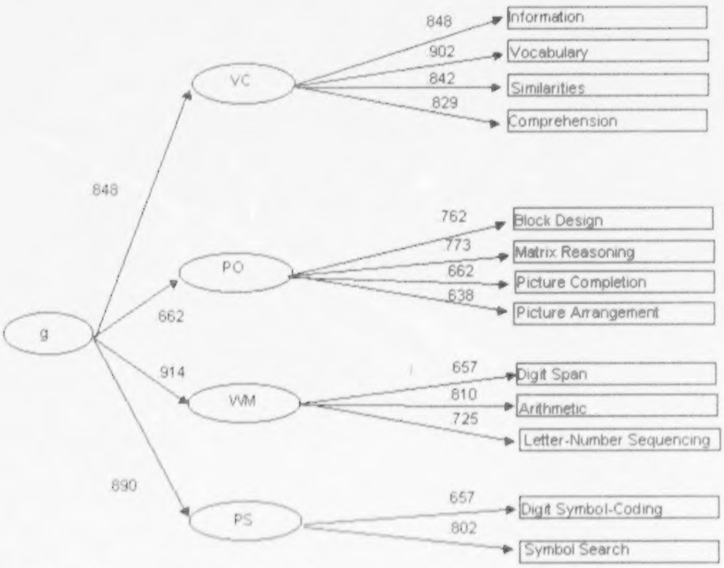

g
.848
VC
.848
Information
.902
Vocabulary
.842
Similarities
.829
Comprehension
.662
PO
.762
Block Design
.773
Matrix Reasoning
.662
Picture Completion
.638
Picture Arrangement
.914
WM
.657
Digit Span
.810
Arithmetic
.725
Letter-Number Sequencing
.890
PS
.657
Digit Symbol-Coding
.802
Symbol Search

Figure 1. Results of a confirmatory factor analysis of the theoretical four factor model of the WAIS-III. VC = Verbal Comprehension, PO = Perceptual Organization, WM = Working Memory, PS = Processing Speed.

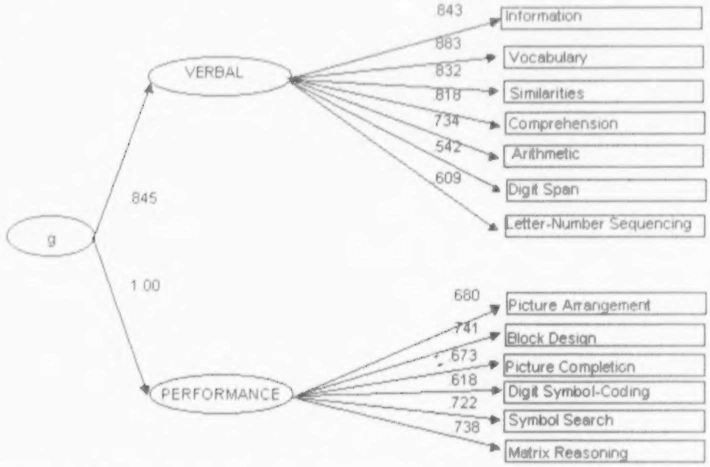

g
.845
VERBAL
.843
Information
.883
Vocabulary
.832
Similarities
.818
Comprehension
.734
Arithmetic
.542
Digit Span
.609
Letter-Number Sequencing
1.00
PERFORMANCE
.680
Picture Arrangement
.741
Block Design
.673
Picture Completion
.618
Digit Symbol-Coding
.722
Symbol Search
.738
Matrix Reasoning

Figure 2. Results of a confirmatory factor analysis of the implied two factor model of the WAIS-III.

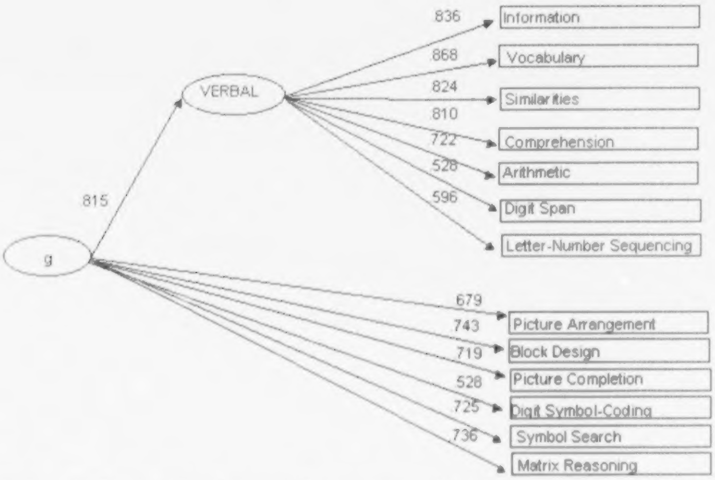

836
Information
868
Vocabulary
VERBAL
824
Similarities
810
722
Comprehension
528
Arithmetic
596
Digit Span
815
Letter-Number Sequencing
g
679
Picture Arrangement
743
Block Design
719
Picture Completion
528
725
Digit Symbol-Coding
736
Symbol Search
Matrix Reasoning

Figure 3. Results of a confirmatory factor analysis of the single intermediate first- order factor and a second-order general factor model of the WAIS-III.

Table 3 presents fit statistics of the three analyses. The first analysis tested the four factor theoretical mode. This analysis produced a χ^2 (df) = 81.79 (61). The TLI (.982) was high and suggested a good fit of the data to the model, as did the other fit indices.

Table 3: Comparisons of Fit Indices of Competing Models of the WAIS-III

Model	χ^2	(df)	AGFI	CFI	TLI	RMSEA	χ^2 diff
1. Four-Factor	81.78	(61)	.907	.986	.982	.041	
2. Implied Two-Factor	147.50	(64)	.830	.943	.930	.081	65.72*
3. Intermediate VIQ	152.92	(65)	.831	.940	.928	.082	71.14*

Note. AGFI = Adjusted goodness of fit index; CFI = Comparison fit index; TLI = Tucker-Lewis Index; RMSEA = Root mean square error of approximation. *$p < .01$.

As can be seen the fit indices of the second and third analyses also suggest a good fit to the data. These additional analyses were conducted to see if the change in parameters would result in an improvement beyond the four factor model. The change in Chi-square and degrees of freedom was used to evaluate the competing models. The results of the second analysis indicate that there was an increase in χ^2 and degrees of freedom (65.72(3), $p < .05$) thus suggesting that the data fit the two factor implied model significantly worse than the four factor model. Similar results were observed in the third analysis for the single first-order intermediate factor model (71.14 (4), $p < .05$). These results suggest the four-factor model provides the most parsimonious fit to the data. Since the Performance factor is subsumed by the second-order g factor, the utility of an implied two factor model was not supported. These results indicate the interpretation of performance on the WAIS-III based on a Verbal /Performance discrepancy or VIQ/PIQ factor structure cannot be supported.

Implication for Practice

Clinicians calculating a Full Scale IQ Index score on the WAIS-III will also obtain a participant's Verbal IQ Index Score, and Performance IQ Index Score. The results of this study indicate that the Performance IQ Index factor is indistinguishable from psychometric g. In other words, psychometric g completely subsumes the Performance factor. This finding is inconsistent with the implied factor structure of the instrument and its construct validity. Therefore, the practitioner is encouraged to exercise caution when making interpretations about first-order VIQ/PIQ differences. Specifically, caution should be employed when identify significant discrepancies between scores on first-order VIQ/PIQ constructs. Since it is necessary to calculate a FSIQ for a participant, it is recommended that interpretation at the first-order factor level be limited to the explicit four factor model.

Summary

This study did not support the VIQ/PIQ dichotomy, which is the hallmark of all Wechsler scales. This particular finding is significant, insofar as construct validity is necessary for treatment validity. "Measurement, even though it is based on observable responses,.would have little meaning or usefulness unless it could be interpreted in light of the underlying theoretical construct" (Crocker & Algina, 1986, p. 7). The first-order PIQ is completely subsumed by the general second-order factor, and therefore the construct validity of a VIQ/PIQ dichotomy was not supported. When compared to the four factor model of the WAIS-III, the two factor and one intermediate factor models did not result in a significant improvement in fit. Therefore, the theoretical structure, rather than the one implied by the calculation of FSIQ appears to be the most parsimonious and accurate portrayal of the WAIS-III's factor structure. In contrast to historical practice, clinicians are not encouraged to make interpretations of a participant's performance on the WAIS-III using a the VIQ/PIQ dichotomy. Rather, they are encouraged to use the four factor model offered by the authors and discussed in detail in the technical manual when making interpretations and recommendations at the first-order factor level.

References

Arbuckle, J.L. (1997). *AMOS users guide version 3.6.* Chicago: SmallWaters.

Bickley, P. G., Keith, T. Z., & Wolfe, L. M. (1995). The three stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, 20, 309-328.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart, & Winston.

Keith, T. Z. (1990). Measurement and design issues in child assessment research. In C. R. Reynolds & R. W. Kamphays (Eds.), *Handbook of psychological and educational assessment of children: Vol 1, Intelligence and achievement* (pp. 29-61). New York: Guilford Press.

Keith, T.Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues (pp. 373-402).* New York: Guilford Press.

Keith, T. A., & Witta, E. L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Quarterly, 12,* 89-107.

The Psychological Corporation. (1997). *Wechsler Adult Intelligence Scale-Third Edition.* San Antonio, TX: Author.

The Psychological Corporation (1997). *WAIS-III WMS-III Technical Manual.* San Antonio, TX: Author.

Sattler, J.M., & Ryan, J.J. (1999). *Assessment of children, Revised: WAIS-III supplement.* LeMesa, CA: Jerome Sattler Publishing.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Weschsler Adult Intelligence Scale-III.* San Antonio, TX: The Psychological Corporation.

Please direct correspondence to:

Dr. Gordon E. Taub
Department of Human Services and Wellness
University of Central Florida
Orlando, FL 32816

(407) 823-0373
e-mail: gtaub@mail.ucf.edu

Descriptors: *Intelligence Tests; *Interpersonal Competence; * Test Interpretation; Performance Factors; Factor Analysis

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Mullane, Jennifer & Stuart J. McKelvie (2001). Effects of removing the time limit on first and second language intelligence test performance. *Practical Assessment, Research & Evaluation*, 7(23). Available online: http://ericae.net/pare/getvn.asp?v=7&n=23. This paper has been viewed 1137 times since 11/6/01.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

▸ Find articles in ERIC written by
Mullane, Jennifer
Stuart J. McKelvie

Effects of Removing the Time Limit on First and Second Language Intelligence Test Performance

Jennifer Mullane and Stuart J. M^c^Kelvie
Department of Psychology
Bishop's University

Abstract

Canadian post-secondary students with a moderate level of second language competence in English or French took the Wonderlic Personnel Test with the standard 12-min time limit or with no time limit. Participants who took the timed test in their second language scored lower than those who took it in their mother tongue, but the disadvantage was greater for limited French-proficient (LFP) students than for limited English-proficient (LEP) students. Scores increased with no time limit and the gain was greater on the French test for LFP students than for mother tongue students. On the English test, the gain was similar for LEP students and mother tongue students. It is concluded that the time accommodation can be applied to clients who are taking an intelligence test in their second language.

Maximum performance psychological tests enjoy widespread use in North America, particularly for educational and employment decisions. Ever since they were administered en masse to U.S. army recruits and immigrants in the early 1900s, the testing of minorities, particularly limited-English proficient (LEP) clients, has been controversial (Samelson, 1977). The issue is test fairness or "intercept bias": the possibility that the minority group scores lower than the majority on the test but not on the criterion (Anastasi & Urbina, 1997). LEP clients may be disadvantaged on a standardized intelligence test given in English, but they may be just as capable in school or on the job as native English speakers (Cummins, 1987).

In the U.S., 15 to 20% of students speak a language other than English at home (Geisenger & Carlson, 1992). In Canada, the corresponding percentage is 32 (23% speaking the second official language, French) (Statistics Canada, 1996). Although it may (Angus, 2000) or may not (Lam, 1993; Rivera, Vincent, Hafner & LaCelle-Peterson, 1997) be appropriate to administer achievement tests to LEP clients in the majority language, doing so with tests of general cognitive functioning (intelligence) is questionable. Test scores of LEP clients will vary with their English skills, and are likely to underestimate their ability to learn, which should remain relatively constant (Angus, 2000). Indeed, even bilingual children perform more poorly than monolinguals on standardized tests (Valdes & Figueroa, 1994).

To alleviate this problem, students could be assessed with a nonverbal test of intelligence, with a test in their own language, or with a modified test in the majority language. Although nonverbal tests have been defended (e.g., Bracken & McCallum, 2001), they may not measure the same cognitive functions as verbal intelligence tests (Angus, 2000; Anastasi & Urbina, 1997), and they are less successful than verbal tests at predicting educational performance (Angus, 2000; Geisenger & Carlson, 1992; Gregory, 2000). Such problems can be avoided by administering the test in the client's own language, but there can be problems in translation, with no guarantee that the second-language version measures the same construct as the original (Wonderlic, 1992). Moreover, few tests have been translated from English (Geisenger & Carlson, 1992).

The third solution is to modify the original test or test conditions to minimize the disadvantage to the LEP client. For example, grammar may be simplified, a glossary of terms may be provided, or time constraints may be relaxed (Azar, 1999; Rivera et al., 1997). In a survey of statewide assessment programs, it was found that 52% of states permitted amendments for LEP students, and the most popular one (81%) was giving extra time. As with test translation, there is no guarantee that such changes will preserve test validity, but an accommodation can be said to work if it provides a "differential advantage" to the LEP client (Azar, 1999). This means that the gain that accrues from the accommodation must be greater for LEP than for English mother tongue clients. It has been found that SAT scores increase for LEP clients who have extra time, but it is not clear if the gain is greater than for other regular students (Azar, 1999). More research is needed on the effect of test modifications (Rivera et al., 1997).

The purpose of the present study is to investigate the effect of removing the time limit for students taking a standardized intelligence test in their second language. To evaluate whether the accommodation worked, the results were compared with those obtained by students who took the test in their mother tongue. In our study, participants completed the 50-item spiral-omnibus Wonderlic Personnel Test (WPT) with the standard time limit or with unlimited time. The WPT is widely employed in personnel selection, and can be used in educational settings (Wonderlic, 1992). Although Murphy (1984) doubts the manual's claim that it measures the "ability to learn," he states that it is a useful predictor of job performance. Furthermore, it can be classified as a test of general intelligence, because the items are based on the Otis Self-administering Tests of Mental Ability, and cover numerical reasoning, verbal reasoning, synonyms-antonyms, nonverbal reasoning, information, and attention to detail (McKelvie, 1992). Moreover, its loadings on Aptitude G of the General Aptitude Test Battery exceed .56, and its scores correlate moderately to very highly (up to .90; Dodrill, 1981; Dodrill & Warner, 1988) with the Wechsler Adult Intelligence Scale in general and psychiatric groups. Wonderlic (1992) also states that the test correlates reasonably (from .30 to .45) with academic achievement, so that it can be used as a selection and counseling tool in postsecondary education. Correlations at the university level are lower than those at high school (McKelvie, 1989, 1992), but this may be due to restriction of range.

Under standard instructions, the WPT is administered with a 12-min time limit, which makes it a highly-speeded test (Davou & McKelvie, 1984). Belcher (1992) states that this may be unfair for the increasing number of LEP clients. When answering an item, they may have to translate it into their mother tongue to fully understand its meaning. By spending more time than the native English speaker on item comprehension, they will not be able to complete as many items in the restricted time allowed. Although they may answer correctly on the items that they attempt, their obtained score will underestimate their true score. Indeed, the Wonderlic (1992) manual reports that Hispanic Americans scored about six points lower than white Americans on the English WPT.

It is also notable that one study has shown that the time accommodation worked on the WPT for students with weak study habits (Davou & McKelvie, 1984). In the standard timed condition, scores were lower for students with weak than with strong study habits. However, the gain with unlimited time was greater for the weak than the strong studiers.

The WPT has also been translated into French for Canadian use, particularly in Québec, where the majority mother tongue is French. Here, English speakers are the minority. On the other hand, in the Québec academic setting where testing took place, the language of instruction is English and students are enrolled from all parts of Canada (as well as the U.S. and overseas). Here, French speakers are the minority. Thus, we had the opportunity to examine the effect of removing the time limit not only for LEP (French mother tongue) clients taking the test in English, but also for LFP (limited-French proficient, English mother tongue) clients taking the test in French. It was hypothesized that, although all participants would benefit from the relaxed time constraint, the gain would be greater for people taking the test in their second language. That is, the gain should be greater on the English WPT for French (LEP) compared to English mother tongue students, and on the French WPT for English (LFP) compared to French mother tongue students.

Method

Participants

The participants were 133 (89 women, 44 men; mean age 21.6 yr.) postsecondary college and university students who reported English or French as their mother tongue and at least moderate competence in the other language. After matching for gender, English and French first-language participants were assigned randomly to one of four experimental conditions resulting from two levels of two independent variables: timing (Limited Time, Unlimited Time) and test language (English, French).

Materials and Procedure

As described above, the Wonderlic Personnel Test (WPT) measures general intelligence. Its reliability has been estimated as .84 to .94 (test-retest, even for a 5-yr, period; Dodrill, 1983), .88 to .94 (split-half; Wonderlic, 1992), and .73 to .95 (alternate-form; Schoenfeldt, 1985).

Although participants reported that they had moderate competence in their second language, this was assessed more formally with two written tests of second-language proficiency: a 13-item self-report measure (Self Evaluation Questionnaire) in their mother tongue, and a 17-item objective test (Objective Cloze Test) in their second language. The Objective Cloze Test (OCT) was developed from an experimental version of the Test of English as a Foreign Language (TOEFL) (Hale, Stansfield, Rock, Hicks, Butler, & Oller, 1989). In the cloze procedure, words in a text are deleted and must be supplied by the examinee. In the OCT, they were given four choices from which to choose. Scores from multiple-choice cloze tests correlate highly with those from fill-in-the-blank versions (Chappelle & Abraham, 1990).

According to Al-Fallay (1997), the cloze procedure has concurrent validity as a predictor of other second-language achievement tests, but its face validity is questionable. We followed their recommendation to supplement the cloze test with another technique, and chose self-assessment, which has been useful (Ross, 1998). The Self Evaluation Questionnaire (SEQ) was adapted from the Bishop's University ESL (English as a Second Language) department's screening questionnaire. Items tapped various aspects of language ability and were answered on a 7-point Likert scale. Examples are: "When I speak French among a small group of people that I know well, I feel 1 (uneasy) to 9 (very much at ease)"; "I can understand newspaper articles without the use of a dictionary: 1 (not at all) to 7 (perfectly)". Both the OCT and SEQ were constructed in English and translated professionally into French.

Participants were tested individually or in small groups. After signing a consent form, they completed the SEQ then the OCT. The SEQ was given first to avoid contamination of self-reports by perceived performance on the objective test. Following these tests, the WPT was administered then participants were debriefed. The WPT was scored using the standard key for the English version. However, for the French version, two adjustments had to be made because of imprecise translation.

Results and Discussion

Second-Language Competence

The correlation between OCT and SEQ scores for all 133 participants was .529, $p < .01$. This indicates that the two tests were related, but tapped slightly different aspects of second-language competence, supporting the view that it should be assessed with more than one technique (Al-Fallay,1997).

In the present design, second-language competence should be moderate and similar in the different experimental conditions. If participants were perfectly bilingual (scoring close to maximum), second-language testing would not an issue, and if they were essentially monolingual (scoring close to 0), they would not be taking a test in another second language. To evaluate this, 2 X 2 X 2 (Timing X Mother Tongue X Test Language) factorial ANOVAs were conducted for each second-language test. There was a significant effect of mother tongue for both OCT scores, $F(1, 125) = 4.49$, $p < .05$, and for SEQ scores, $F(1, 125) = 11.57$, $p < .01$. Mother-tongue French speakers scored higher on the English OCT than mother-tongue English speakers scored on the French OCT (Ms = 12.3. 11.3). French speakers also rated themselves as more proficient in English than English speakers rated themselves in French (Ms = 62.5, 53.6). Although the tests were professionally translated from English to French, the OCT may not be equally difficult in each language. However, the fact that the results agreed with those of the SEQ indicates that our French speakers were more proficient in English than were our English speakers in French. At the same time, none of the mean scores was extreme. On the OCT, the maximum possible score was 17 (Ms were 12.3, 11.3) and on the SEQ it was 91 (Ms were 62.5, 53.6). Moreover, the lack of any significant interactions indicates that second-language competence was matched in the four timing/test language conditions.

Because the major comparisons of interest were the effects of time for people taking the test in their first or second language, the different levels of second-language competence were dealt with by including OCT and SEQ scores as covariates in the analysis of WPT scores.

Intelligence Test Performance

Initially, a 2 X 2 X 2 (Timing X Mother Tongue X Test Language) factorial ANOVA (with SEQ and OCT as covariates) was conducted on WPT scores. Table 1 shows the means in each of the eight conditions. Not surprisingly, there was a significant effect of timing, $F(1, 123) = 150.46$, $p < .001$, with higher scores in the unlimited time than in the limited time condition (Ms = 32.58, 22.17). Converted to Cohen's (1977) standardized effect size (d), this difference was 2.10 which clearly exceeds his guideline of 0.80 for a large effect.

Table 1: Mean Wonderlic Personnel Test Scores in Each Condition

	Limited Time			Unlimited Time		
Test Language	n	M	SD	n	M	SD
English						
English Mother Tongue	16	26.33	4.24	17	34.24	6.08
French Mother Tongue (LEP)	18	23.03	5.86	13	30.70	7.50
French						
English Mother Tongue (LFP)	17	16.46	4.62	16	31.94	5.10
French Mother Tongue	18	22.74	4.36	18	33.49	3.07

Note. Maximum score = 50. Means are adjusted for covariates. LEP = limited-English proficient, LFP = limited-French proficient.

There was also a significant effect of test language, $F(1, 123) = 9.32$, $p < .01$, with lower scores on the French WPT than on the English WPT ($Ms = 26.09, 28.66$). However, timing interacted significantly with test language, $F(1, 123) = 12.15$, $p < .01$. The gain from the limited to the unlimited time condition was greater on the French test ($Ms = 19.42, 32.76$; gain = 13.34) than on the English test ($Ms = 24.92, 32.39$; gain = 7.47). There was also a significant interaction between mother tongue and test language, $F(1, 123) = 13.92$, $p < .01$. For people whose mother tongue was English, scores were lower on the French than on the English test ($Ms = 24.29, 30.03$), but for those whose mother tongue was French, the scores were very similar ($Ms = 27.88, 27.28$).

These results indicate that the effects of timing and of mother tongue were greater on the French than on the English test. Because it was predicted that the effect of timing would be greater for second- than for first-language participants on each test, planned 2 X 2 (Timing X Mother Tongue) ANOVAs (again with OCT and SEQ as covariates) were conducted separately for the English and French WPTs. For the French version of the test, all three effects were significant: timing, $F(1, 63) = 157.18$, $p < .01$, mother tongue, $F(1, 63) = 13.15$, $p < .01$, and their interaction, $F(1, 63) = 5.04$, $p < .05$. Scores were higher in the unlimited than in the limited time condition, and for French mother tongue than for English mother tongue (LFP) participants. However, and of particular interest here, the latter difference was only significant with limited time. Here, LFP students (English native speakers) performed more poorly than French native speakers, $t(63) = 4.24$, $p < .01$, but when the time constraint was removed, they did not, $t(63) = 1.07$, $p > .10$. Another way of looking at this is that both groups of participants benefited from extra time, but the gain was greater for the LFP students (15.5 points, $d = 3.54$) than for the French native speakers (10.8 points, $d = 2.44$). In other words, the time accommodation worked because it provided a differential advantage to the LFP clients (Azar, 1999).

For the English version of the test, there were significant effect of timing, $F(1, 58) = 34.96$, $p < .01$, and of mother tongue, $F(1, 58) = 6.31$, $p < .05$, but *not* their interaction, $F(1, 58) = 0.01$, $p > .90$. Scores were higher in the unlimited than in the limited time condition, and for English mother tongue than for French mother tongue (LEP) participants. Thus, although LEP students performed more poorly than English native speakers with the standard 12-min time limit, and although they benefited from unlimited time, the gain was not greater than that for English speakers. In fact, the effect size for time was $d = 1.46$ (7.7 points) for French speakers and $d = 1.50$ for English speakers (7.9 points). Here, the time accommodation did not work.

Why did the time accommodation work for LFP but not for LEP participants? The answer may be that, in the timed condition, the LFP disadvantage on the French test ($Ms = 16.46, 22.74$; difference = 6.28) was greater than the LEP disadvantage on the English test ($Ms = 23.03, 26.33$; difference = 3.30). The accommodation may only be effective if the disadvantage is great.

But why was the LFP disadvantage greater than the LEP disadvantage? The most obvious reason is that French second-language competence was less than English second-language competence as assessed on the OCT and SEQ. In fact, the mean levels of competence were not extreme, and the differences were controlled via covariance analysis[1]. Perhaps the answer is that the tests did not fully capture the fact that the participants whose second language was English were studying in an English-speaking institution, and who probably had more practice listening, reading and writing in their second language than did participants who second language was French. In fact, the SEQ only had one question about frequency of second-language use, and it only referred to speaking. To aid in the identification of clients likely to perform poorly on the timed WPT in their second language, the SEQ should be expanded to include information about reading and writing. It might also be noted that if the present study had also been conducted in a French-speaking institution, the results might have been reversed. That is, the gain from unlimited time might have been greater for LFP than for LEP students.

Although these results show that the time accommodation can work, they do not show whether the WPT scores obtained with unlimited time are as valid as those obtained with limited time. Wonderlic (1992) himself discusses this issue, stating that "while untimed scores are valid assessments of cognitive ability, they are not as accurate as the timed scores." Notably, a 25-item short form of the WPT given with unlimited time was as reliable as the full version (when corrected for length), and also had a similar criterion validity coefficient for predicting university grades (McKelvie, 1994). However, because his studies indicated that people taking the test under both conditions scored about six points higher with unlimited time, Wonderlic recommends that this score be used to estimate the timed score by subtracting six points from it.

French mother-tongue speakers (LEP participants) scored only slightly lower than English mother-tongue speakers on the English WPT with unlimited time. Because English mother-tongue speakers (LFP) did not score significantly lower than French mother-tongue speakers on the French WPT in this condition, extra time minimized or removed the second language disadvantage. Therefore, we suggest that people taking an intelligence test in their second language be permitted the accommodation of unlimited time. In the case of the WPT, their timed score can then be estimated by subtracting six points.

Conclusion

These results provide experimental evidence that the time accommodation can work for people whose second-language intelligence test limited time score is clearly lower than that of mother-tongue participants, and it does no injustice to those whose limited time score is only slightly lower. Therefore, we recommend removing time limits on standardized intelligence tests for clients taking them in their second language. The present measures of second-language competence should be expanded, and future research should obtain more information about the psychometric properties (norms, reliability, validity) of untimed intelligence tests.

-

References

Al-Fallay, I. (1997). Investigating the reliability and validity of the fixed ratio multiple-choice cloze test. *Human and Social Sciences, 24*, 507-526.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*, 7th ed. Upper Saddle River, NJ: Prentice-Hall.

Angus, W. A. (2000). Using achievement tests, diagnostic (achievement) tests, and tests of intelligence with ESP populations. http://www.psychtest.com/ESLtest.htm.

Azar, B. (1999). Fairness a challenge when developing special needs tests. *APA Monitor Online, 30*. http://www.apa.org/monitor/dec99/in2.html.

Belcher, M. J. (1992). Review of the Wonderlic Personnel Test. In J. J. Kramer & J. C. Conoley (eds.), *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Bracken, B. & M^{c}Callum, R. S. (2001). Assessing intelligence in a population that speaks more than tow hundred languages: A nonverbal solution. In L. A. Suzuki and J. G. Ponterotto (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications*, 2nd ed. San Francisco, CA: Jossey-Bass, Inc.

Chapelle, C. A., & Abraham, R. G. (1990)/ Cloze method: What difference does it make? *Language Testing, 7*, 121-146.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Cummins, J. (1997). Psychoeducational assessment in multicultural school systems. *Canadian Journal for Exceptional Children, 3*, 115-117.

Davou, D., & McKelvie, S. J. (1984). Relationship between study habits and performance on an intelligence test with limited and unlimited time. *Psychological Reports, 54*, 367, 371.

Dodrill, C. B. (1981). An economical method for the evaluation of general intelligence in adults. *Journal of Consulting and Clinical Psychology, 49*, 668-673.

Dodrill, C. B. (1983). Long-term reliability of the Wonderlic Personnel Test. *Journal of Consulting and Clinical Psychology, 51*, 316-327.

Dodrill, C. B., & Warner, M. H. (1988). Further studies of the Wonderlic Personnel Test as a brief measure of intelligence. *Journal of Consulting and Clinical Psychology, 56*, 145-147.

Edinger, J. D., Shipley, R. H., Watkins, C. E., & Hammett, E. B. (1985). Validity of the Wonderlic Personnel Test as a brief IQ measure in psychiatric patients. *Journal of Consulting and Clinical Psychology, 53*, 937-939.

Geisenger, K. F., & Carlson, J. F. (1992). Assessing language-minority students. Practical Assessment, Research & Evaluation, 3(2). Available online: http://ericae.net/pare/getvn.asp?v=3&n=2.

Gregory, R. J. (2000). *Psychological testing: History, principles, and applications*, 3rd ed. Boston: Allyn and Bacon.

Hale, G., Stansfield, C., Rock, D., Hicks, M., Butler, F., & Oller, J. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6*, 49-78.

Lam, T. C. M. (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26*, 179-191.

McKelvie, S. J. (1989). The Wonderlic Personnel Test: Reliability and validity in an academic setting. *Psychological Reports, 65*, 161-162.

McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? *The Journal of Psychology, 119*, 59-72.

McKelvie, S. J. (1994). Validity and reliability findings for an experimental short form of the Wonderlic Personnel Test in an academic setting. *Psychological Reports, 75*, 907-910.

Murphy, K. R. (1984). The Wonderlic Personnel Test. In D. J. Keyser and R. C. Sweetland (Eds.), *Test critiques* (volumes I-VI). Kansas City, MO: Test Corporation of America, pp. 769-775.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). Statewide assessment programs: Policies and practicies for the inclusion of limited English proficient students. *Practical Assessment, Research & Evaluation, 5(13)*. Available online: http://ericae.net/parc/getvn.asp?v=3&n=2.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiental factors. *Language Testing, 15*, 1-20.

Samelson, F. (1977). World War I intelligence testing and the development of psychology. *Journal of the History of the Behavioral Sciences, 13*, 274-282.

Schoenfeldt, L. F. (1985). Review of the Wonderlic Personnel Test, In J. V. Mitchell Jr. (ed.). *The ninth mental measurements yearbook.* Vol 2. Lincoln, NE: University of Nebraska Press.

Statistics Canada (1996). 1996 Census figures. http://www.statcan.ca/english/Pgdb/People/Population/demo29d.htm).

Valdes, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Stamford, CT: Ablex Publishing Corporation.

Wonderlic, E. F. (1992). *Wonderlic Personnel Test User's Manual.* Libertyville, IL: E. F. Wonderlic.

Footnote

[1] A second analysis of WPT scores was conducted in which the sample size was reduced to 109 by removing participants with higher English and lower French second-language OCT and SEQ scores. The results were the same, with the exception that LEP participants did not score significantly lower than English mother tongue participants on the English WPT. The key point is that, once again, the gain in scores from extra time was only of significant benefit to LFP participants on the French WPT.

Author Note

We thank Dr. Richard Kruk for advice on second language proficiency testing, Denise Bernier for pointing out translation ambiguities, and four reviewers for helpful critical comments. Send correspondence to Stuart J. McKelvie, Department of Psychology, Bishop's University, Lennoxville, Québec J1M 1Z7, Canada. Electronic address is smckelvi@ubishops.ca.

Descriptors: Bilingual Education; Test Format; Evaluation Methods; Intelligence Tests; Language Proficiency; Time Factors [Learning]

Home Articles Subscribe Review Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ Go

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Kellow, J. Thomas & Victor L. Willson (2001). Consequences of (mis)use of the texas assessment of academic skills (taas) for high-stakes decisions: a comment on haney and the texas miracle in education. *Practical Assessment, Research & Evaluation*, 7(24). Available online: http://ericae.net/pare/getvn.asp?v=7&n=24. This paper has been viewed 1048 times since 12/7/01.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

▸ Find articles in ERIC written by
Kellow, J. Thomas
Victor L. Willson

Consequences of (mis)use of the Texas Assessment of Academic Skills (TAAS) for high-stakes decisions: A comment on Haney and the Texas miracle in education.

J. Thomas Kellow, University of Houston
Victor L. Willson, Texas A&M University

Abstract

This brief paper explores the consequences of failing to incorporate measurement error in the development of cut-scores in criterion-referenced measures. The authors use the case of Texas and the Texas Assessment of Academic Skills (TAAS) to illustrate the impact of measurement error on "false negative" decisions. These results serve as further evidence to support Haney's (2000) contentions regarding the (mis)use of high-stakes testing in the state of Texas.

Walt Haney's (2000) treatise on *The Myth of the Texas Miracle in Education* highlights a number of concerns related to high-stakes decision-making in K-12 settings. His elaboration of the history and development of the Texas Assessment of Academic Skills (TAAS) was illuminating, especially for those unfamiliar with the often capricious fashion in which standards are ultimately set for high-stakes decisions. The evidence Haney presents in evaluating the TAAS fits well with Samuel Messick's (1994) argument for considering the *consequences* of test use. Although sometimes referred to as *consequential validity*, Messick considered this aspect of test interpretation and use to be one of many forms of *construct validity*. His view of the evaluative role of validity is summarized nicely in the following paragraph:

> When assessing any of these constructs – whether attributes of persons or groups, of objects or situations – validity needs to be systematically addressed, as do other basic measurement issues such as reliability, comparability, and fairness. This is so because validity, comparability, and fairness are not just measurement principles, they are *social values* that have meaning and force outside of measurement whenever evaluative judgment and decisions are made. As such, validity assumes both a scientific and political role that can by no means be fulfilled by a simple correlation coefficient between test scores and a purported criterion or by expert judgments that test content is relevant to the proposed test use (p. 1).

One technical aspect of the TAAS that Haney takes to task is the reliability of scores yielded by the test and the potential for misclassification as a function of measurement error. As the author notes, "The reason the setting of passing scores on a high-stakes test such as the TAAS is so important is that the passing score divides a continuum of scores into just two categories, pass and fail. Doing so is hazardous because all standardized test scores contain some degree of measurement error" (Haney 2000, p. 10). Haney also points out that measures of test-retest (or alternate-form) reliability and not internal consistency should ideally be used to inform judgment as to the potential for misclassfication due to measurement error. In light of the conspicuous absence of test-retest data on the TAAS, he attempts to use extant data to approximate this reliability estimate as compared to the internal consistency (KR20) estimates provided by the Texas Education Agency (TEA). Although the approach he uses is somewhat problematic (see Wainer, 1999), it is clear that these test-retest estimates are lower than the KR20 estimates.

The thrust of Haney's argument is that measurement error inherent in the TAAS (or any other measure, for that matter) contributes appreciably to the rate of "false negatives," or misclassifying *passing* students as *non-passing*. His focus, however, is exclusively on the tenth-grade Exit-Level TAAS since this is the test high school students in the state of Texas must pass in order to graduate from high school (although it is not the sole criterion). But in some districts, the TAAS is also being used at other grade levels for high-stakes decisions – namely, promotion and retention of students. Waco Independent School District and Houston Independent School District, among others, require students in grades 3 through 8 to pass all portions of the TAAS in order to be considered for promotion to grade level. Students who fail any subtest during the statewide spring administration can expect to attend mandatory summer school, after which they are tested on a "released" version of the TAAS. Those students who fail are held back in grade. This stands in contrast to the Exit-Level testing schedule where, as Haney notes, students have as many as eight opportunities to pass the exam. In addition, the TEA has announced plans to use the TAAS for promotion/retention decisions in the third, fifth, and eighth grades beginning in the fall of 2001 (Texas Education Agency, 1999).

Given the current political climate and the clarion calls for school district accountability throughout the nation, the number of districts in Texas using the TAAS for promotion and retention decisions in all tested grades no doubt will increase. Indeed, the Houston model has been lauded for its strict criteria for student grade promotion, which is believed to increase student motivation and achievement while *reducing* the student dropout rate (Markley, 1999). Our purpose in this response is to elaborate on the potential social consequences of using the TAAS for high-stakes decisions, specifically grade promotion and retention. The question we asked is rather simple: "Given the imperfect reliability of the TAAS test, how many students with a true passing score are potentially misclassified as failing across grades and subtests in a given year"? Fortunately, sufficient summary statistics and frequency distributions were available from the TEA to estimate these numbers.

Method

The most recent reliability estimates for the TAAS were reported for the 1998-99 school year; therefore we used means, standard deviations, and frequency distributions for this same year. We want to emphasize that there are a number of ways to go about this estimation process. The method presented emerged because it is fairly intuitive and required minimal summary statistics from the data.

First, we should note that the 70% passing standard on the TAAS mentioned by Haney is *not* fixed across grades and subtests. Because of differential item and thus test difficulty, the proportion of items correct needed to pass a given subtest ranges from .64 to .75, according to the TEA. The Texas Learning Index (TLI) was developed by the TEA for the purpose of, among other things, providing a consistent passing standard across test forms. The TLI is a linear standardized scoring transformation with a standard deviation of 15 and an anchor (rather than mean) of 70, which represents the passing standard for a given subtest at a given grade. The TLI is calculated in z-score form as:

$$\mathrm{TLI} = [(z_{\mathrm{observed}} - z_{\mathrm{passing}})*15] + 70 \qquad (1)$$

Although the TLI observed score of 70 is the passing standard, this standard fails to incorporate measurement error in determining the appropriate cut score. Specifically, the process of modifying cut-scores involves determining the domain score in proportion-correct form that constitutes mastery (τ_o) and then adjusting this value to estimate a new cut score (X_0) in number-correct form that reflects the measurement error in the test data (Crocker & Algina, 1986). Huynh and Sanders (1980) provide an approximate procedure for this purpose that works well when a test consists of 20 or more items and the observed proportion correct cut score falls within .50 to .80. The TAAS subtests meet both criteria. This formula is given as:

$$X_0 = \frac{n - KR_{21}}{KR_{21}} \tau_0 + \frac{KR_{21} - 1}{KR_{21}} \mu_x + .5 \qquad (2)$$

Where *n* is the number of items on the test, τ_0 is the observed proportion-correct cut score, and μ_x is the mean number of items correct. As noted by Crocker and Algina (1986), as KR21 approaches 1.0, X_0 approaches $n\ \tau_o + .5$ irrespective of the value of μ_x.

Our question focuses on the estimate X_0 when transformed to a TLI score metric. Put simply: What is the passing TLI adjusted for measurement error for a given subtest in a given grade, and what percent and number of students met this adjusted criterion but not the standard cut score of 70? The following steps were employed to determine X_0 in TLI form:

1. calculate X_0 in raw score form;
2. transform X_0 into a z-score;
3. substitute z X_0 for z observed in Formula 1.

Results

Table 1 provides the adjusted cut score X_0 in the TLI metric across grades for both the mathematics and reading subtests.

Table 1
TLI passing cut-scores adjusted for measurement error by subtest and grade

Grade	*Reading*	*Mathematics*
3rd	67.5	67.4
4th	67.2	67.6
5th	67.2	67.0
6th	67.6	68.2
7th	67.9	68.4
8th	67.7	68.3
10th	66.9	68.7

We then used TLI frequency distributions for the 1998-99 administration of the TAAS to determine the percent and number of students who received a TLI score of X_0 or higher but less that 70. Because the TLI frequency distribution tables obtained from TEA report whole number values, we rounded the obtained X_0 estimates to the closest whole number. These data are presented in Table 2 disaggregated by subtest and grade.

Table 2
Percent and number of potentially misclassified students by subtest and grade

Grade	*Reading*	*Mathematics*
3rd	1.7 (n=4,243)	2.9 (n=7,151)
4th	1.7 (n=4,105)	2.1 (n=5,239)
5th	2.2 (n=5,509)	2.4 (n=6,007)
6th	2.2 (n=5,725)	2.5 (n=6,653)
7th	2.0 (n=5,410)	2.8 (n=7,474)
8th	1.4 (n=3,620)	2.6 (n=6,903)
10th	2.9 (n=6,570)	1.6 (n=3,650)

Because of the rounding procedure mentioned earlier, these percentages and student numbers are approximations. Roughly 35,182 students who took the reading subtest in the 1998-99 school year were classified as failing, despite having an observed score that would have met (or exceeded) the passing criterion had the presence of measurement error been incorporated into the cut score. On the mathematics subtest, 43,077 students who failed met (or exceeded) the adjusted observed criterion score.

Discussion

Because all tests are inherently unreliable to some degree, measurement errors must be accommodated in the development of cut-scores for criterion-referenced tests, particularly when these instruments are used to make high-stakes decisions for student placement. Our analysis focused exclusively on the impact of false negative classification errors. There exists, of course, a second type of misclassification termed a "false positive," or misclassifying *non-passing* students as *passing*. Although both types of misclassification are serious, a survey of Texas educators conducted by Haney (2000) as part of his investigation of the TAAS indicated that respondents viewed the consequences of denying a high school diploma to a qualified student based on a classification error (false negative) as considerably more serious that granting a diploma to an unqualified student (false positive). Indeed, only the consequences to society of granting a license to an unqualified pilot, physician, or teacher, respectively, were viewed as more serious. Additionally, the literature on grade retention is fairly consistent in noting the deleterious effects of these policies (e.g., increased dropout rates), particularly when strong individualized remediation procedures are not in place (McCoy & Reynolds, 1999). Put simply, based on Haney's (2000) survey and the empirical findings on the consequences of retaining students, we feel it seems reasonable to place greater emphasis on the occurrence of false negatives -- at least in the context of education and student promotion decisions.

The estimation process we employed produced results suggesting that about 2% of students who take the state-mandated TAAS exam will be scored as false negatives on one or more of the subtests. The consequences of misclassification will become more evident as the TAAS is increasingly used for promotion and retention decisions in Texas. There is, however, a much larger picture emerging at the national level regarding the (mis)use of standardized assessment tools. To put this in a broader perspective, we extended our analysis to include an estimate of how many students nationally would be misclassified as false negatives if a testing program such as the TAAS were in place. We assumed that testing would include the same grade levels as the Texas accountability model, and assumed also a national testing instrument with the same technical adequacy (reliability) as the TAAS. National data were obtained for the 1998-99 school year disaggregated by grade level. Combining both reading and mathematics error rates results in approximately *1.1 million students* that would potentially be misclassified as false negatives each year across the country. Two percent clearly is no small number in the national context.

The recently installed Bush administration has issued school accountability reform measures that rely almost exclusively on standardized achievement tests. Many questions remain, however, regarding the structure and implementation of the testing program that will serve as a measure of student performance across states. What these tests will look like, the extent to which they yield scores that are psychometrically meaningful, and the importance of the scores in guiding student-level decisions are issues that have not been addressed to date. It is notable that both the American Psychological Association (APA) and, more recently, the American Educational Research Association (AERA) have issued position statements advising against the use of a single assessment for high-stakes decisions at the individual level. It seems probable, however, that the national appetite for school accountability in the form of student achievement scores will overwhelm any concerns over the ethical consequences of high-stakes testing.

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando: Harcourt Brace.

Haney, W. (2000). The myth of the Texas Miracle in education. *Education Policy Analysis Archives [On-line serial], 8* (41). Available: *http://epaa.asu.edu/epaa/v8n41/.*

Huynh, H., & Saunders, J. C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement, 17,* 351-358.

Markley, M. . (1999, September 8). HISD rules holding more students back: Expanded standards cut social promotions. *Houston Chronicle,* p. A1.

Messick, S. A. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment, 10,* 1-9.

McCoy, A. R., & Reynolds, A. J. (1999). Grade retention and school performance: An extended investigation. *Journal of School Psychology, 37*, 273-298.

Texas Education Agency (1999, June 22). Briefing book: Legislation affecting public education [On-line]. Available: http://www.tea.state.tx.us/brief/doc2.html.

Wainer, H. (1999). Comments on the ad hoc committee's critique of the Massachusetts Teacher Tests. *Education Policy Analysis Archives [On-line serial], 7* (5). Available: *http://epaa.asu.edu/epaa/v7n5.html.*

Descriptors: Test Validity; Test reliability; Consequences; Impact; Error

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). Available online: http://ericae.net/pare/getvn.asp?v=7&n=25. This paper has been viewed 2414 times since 12/11/01.

▸ Find similar papers in
ERICAE Full Text Librar
Pract Assess, Res & Ev
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Pr

▸ Find articles in ERIC writte
Mertler, Craig A.

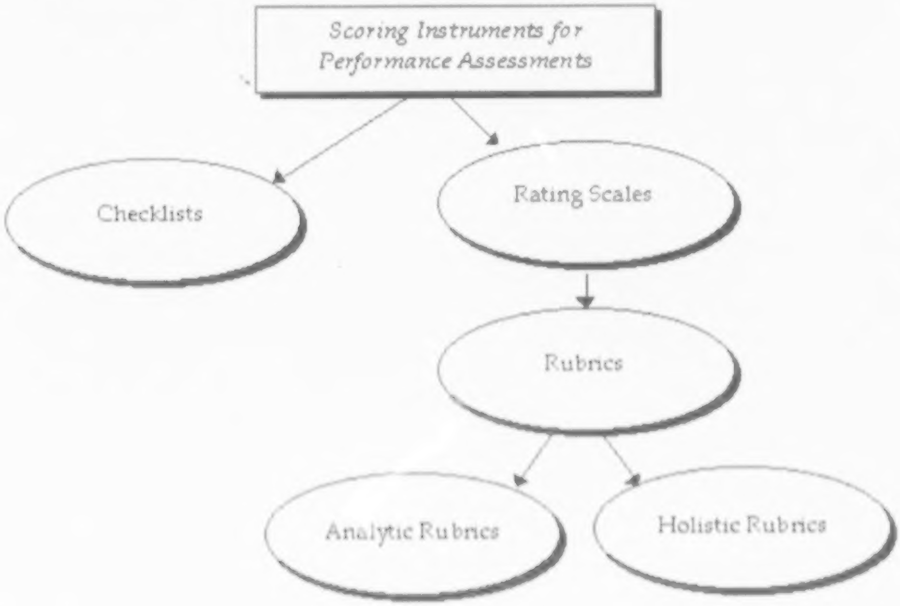
Designing Scoring Rubrics for Your Classroom

Craig A. Mertler
Bowling Green State University

Rubrics are rating scales-as opposed to checklists-that are used with performance assessments. They are formally defined as scoring guides, consisting of specific pre-established performance criteria, used in evaluating student work on performance assessments. Rubrics are typically the specific form of scoring instrument used when evaluating student performances or products resulting from a performance task.

There are two types of rubrics: holistic and analytic (see Figure 1). A **holistic rubric** requires the teache score the overall process or product as a whole, without judging the component parts separately (Nitko, In contrast, with an **analytic rubric**, the teacher scores separate, individual parts of the product or perfo first, then sums the individual scores to obtain a total score (Moskal, 2000; Nitko, 2001).

Figure 1:
Types of scoring instruments for performance assessments



Holistic rubrics are customarily utilized when errors in some part of the process can be tolerated provide overall quality is high (Chase, 1999). Nitko (2001) further states that use of holistic rubrics is probably appropriate when performance tasks require students to create some sort of response and where there is definitive correct answer. The focus of a score reported using a holistic rubric is on the overall quality, proficiency, or understanding of the specific content and skills-it involves assessment on a unidimensio (Mertler, 2001). Use of holistic rubrics can result in a somewhat quicker scoring process than use of ana rubrics (Nitko, 2001). This is basically due to the fact that the teacher is required to read through or othe examine the student product or performance only once, in order to get an "overall" sense of what the stu was able to accomplish (Mertler, 2001). Since assessment of the overall performance is the key, holistic are also typically, though not exclusively, used when the purpose of the performance assessment is sum in nature. At most, only limited feedback is provided to the student as a result of scoring performance ta this manner. A template for holistic scoring rubrics is presented in Table 1.

Table 1:
Template for Holistic Rubrics

Score	Description
5	Demonstrates complete understanding of the problem. All requirements of task are included in response.
4	Demonstrates considerable understanding of the problem. All requirements of task are included.
3	Demonstrates partial understanding of the problem. Most requirements of task are included.
2	Demonstrates little understanding of the problem. Many requirements of task are missing.
1	Demonstrates no understanding of the problem.
0	No response/task not attempted.

Analytic rubrics are usually preferred when a fairly focused type of response is required (Nitko, 2001); t for performance tasks in which there may be one or two acceptable responses and creativity is not an es feature of the students' responses. Furthermore, analytic rubrics result initially in several scores, followe summed total score-their use represents assessment on a multidimensional level (Mertler, 2001). As pre mentioned, the use of analytic rubrics can cause the scoring process to be substantially slower, mainly b assessing several different skills or characteristics individually requires a teacher to examine the product times. Both their construction and use can be quite time-consuming. A general rule of thumb is that an individual's work should be examined a separate time for each of the specific performance tasks or scori criteria (Mertler, 2001). However, the advantage to the use of analytic rubrics is quite substantial. The d feedback offered to students-and to teachers-is significant. Students receive specific feedback on their performance with respect to each of the individual scoring criteria-something that does not happen whe holistic rubrics (Nitko, 2001). It is possible to then create a "profile" of specific student strengths and weaknesses (Mertler, 2001). A template for analytic scoring rubrics is presented in Table 2.

Table 2:
Template for analytic rubrics

	Beginning **1**	**Developing** **2**	**Accomplished** **3**	**Exemplary** **4**	**Score**
Criteria #1	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #2	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #3	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #4	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	

Prior to designing a specific rubric, a teacher must decide whether the performance or product will be sc holistically or analytically (Airasian, 2000 & 2001). Regardless of which type of rubric is selected, spec performance criteria and observable indicators must be identified as an initial step to development. The regarding the use of a holistic or analytic approach to scoring has several possible implications. The mo important of these is that teachers must consider first how they intend to use the results. If an overall, summative score is desired, a holistic scoring approach would be more desirable. In contrast, if formativ feedback is the goal, an analytic scoring rubric should be used. It is important to note that one type of ru not inherently better than the other-you must find a format that works best for your purposes (Montgom 2001). Other implications include the time requirements, the nature of the task itself, and the specific performance criteria being observed.

As you saw demonstrated in the templates (Tables 1 and 2), the various levels of student performance c defined using either quantitative (i.e., numerical) or qualitative (i.e., descriptive) labels. In some instanc teachers might want to utilize both quantitative and qualitative labels. If a rubric contains four levels of proficiency or understanding on a continuum, quantitative labels would typically range from "1" to "4." using qualitative labels, teachers have much more flexibility, and can be more creative. A common type qualitative scale might include the following labels: master, expert, apprentice, and novice. Nearly any t qualitative scale will suffice, provided it "fits" with the task.

One potentially frustrating aspect of scoring student work with rubrics is the issue of somehow converti to "grades." It is not a good idea to think of rubrics in terms of percentages (Trice, 2000). For example, i rubric has six levels (or "points"), a score of 3 should not be equated to 50% (an "F" in most letter gradi systems). The process of converting rubric scores to grades or categories is more a process of logic than

BEST COPY AVAILABLE

mathematical one. Trice (2000) suggests that in a rubric scoring system, there are typically more scores average and above average categories (i.e., equating to grades of "C" or better) than there are below ave categories. For instance, if a rubric consisted of nine score categories, the equivalent grades and categori might look like this:

Table 3:
Sample grades and categories

Rubric Score	*Grade*	*Category*
8	A+	Excellent
7	A	Excellent
6	B+	Good
5	B	Good
4	C+	Fair
3	C	Fair
2	U	Unsatisfactory
1	U	Unsatisfactory
0	U	Unsatisfactory

When converting rubric scores to grades (typical at the secondary level) or descriptive feedback (typical elementary level), it is important to remember that there is not necessarily one correct way to accomplis The bottom line for classroom teachers is that they must find a system of conversion that works for the fits comfortably into their individual system of reporting student performance.

Steps in the Design of Scoring Rubrics

A step-by-step process for designing scoring rubrics for classroom use is presented below. Information f procedures was compiled from various sources (Airasian, 2000 & 2001; Mertler, 2001; Montgomery, 2 Nitko, 2001; Tombari & Borich, 1999). The steps will be summarized and discussed, followed by prese of two sample scoring rubrics.

Step 1: *Re-examine the learning objectives to be addressed by the task*. This allows you match your scoring guide with your objectives and actual instructi

Step 2: *Identify specific observable attributes that you want to see (as well as those you don't want to see) your students demonstrate in their product, process, or performance*. Specify the characteristics, skills, or behaviors that you will be looking for, as well as common mistakes you do not want to see.

Step 3: *Brainstorm characteristics that describe each attribute*. Identify ways to describe abov average, average, and below average performance for each observable attribute identifi in Step 2.

Step 4a: *For holistic rubrics, write thorough narrative descriptions for excellent work and poor work incorporating each attribute into the description*. Describe the highest and lowes levels of performance combining the descriptors for all attributes.

Step 4b: *For analytic rubrics, write thorough narrative descriptions for excellent work and poo work for each individual attribute*. Describe the highest and lowest levels of performance using the descriptors for each attribute separately.

Step 5a: *For holistic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for the collective attributes*. Write descriptions for all intermediate levels of performance.

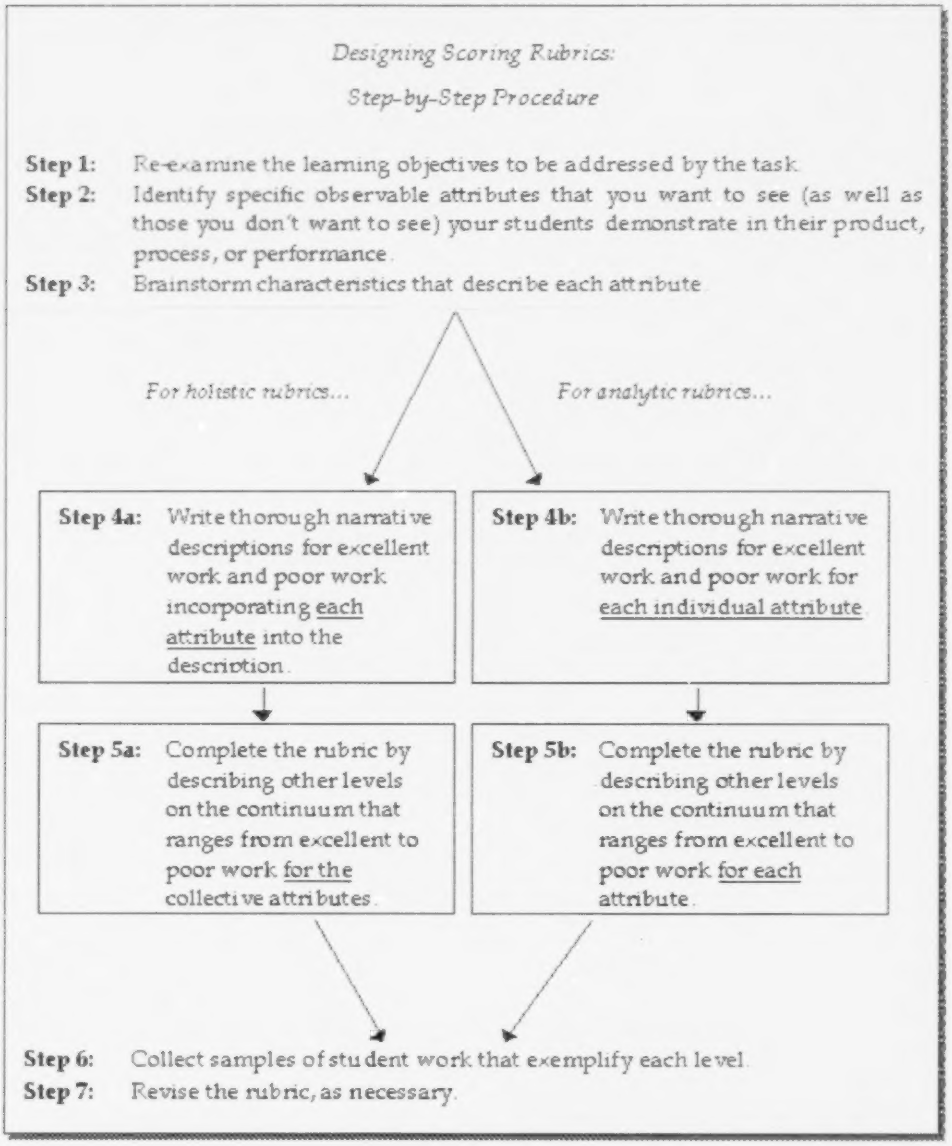
Step 5b: *For analytic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for each attribute*. Write descriptions for all intermediate levels of performance for each attribute separately.

Step 6: *Collect samples of student work that exemplify each level*. These will help you score in the future by serving as benchmarks.

Step 7: *Revise the rubric, as necessary*. Be prepared to reflect on the effectiveness of the rubri and revise it prior to its next implementation.

These steps involved in the design of rubrics have been summarized in Figure 2.

Figure 2:
Designing Scoring Rubrics: Step-by-step procedures



Designing Scoring Rubrics:

Step-by-Step Procedure

Step 1: Re-examine the learning objectives to be addressed by the task.

Step 2: Identify specific observable attributes that you want to see (as well as those you don't want to see) your students demonstrate in their product, process, or performance.

Step 3: Brainstorm characteristics that describe each attribute.

For holistic rubrics...

For analytic rubrics...

Step 4a: Write thorough narrative descriptions for excellent work and poor work incorporating each attribute into the description.

Step 4b: Write thorough narrative descriptions for excellent work and poor work for each individual attribute.

Step 5a: Complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for the collective attributes.

Step 5b: Complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for each attribute.

Step 6: Collect samples of student work that exemplify each level.

Step 7: Revise the rubric, as necessary.

Two Examples

Two sample scoring rubrics corresponding to specific performance assessment tasks are presented next. discussions precede the actual rubrics. For illustrative purposes, a holistic rubric is presented for the firs and an analytic rubric for the second. It should be noted that either a holistic or an analytic rubric could been designed for either task.

Example 1:
Subject - Mathematics

Grade Level(s) - Upper Elementary

Mr. Harris, a fourth-grade teacher, is planning a unit on the topic of data analysis, focusing primarily on skills of estimation and interpretation of graphs. Specifically, at the end of this unit, he wants to be able assess his students' mastery of the following instructional objectives:

- Students will properly interpret a bar graph.
- Students will accurately estimate values from within a bar graph. (step 1)

Since the purpose of his performance task is summative in nature - the results will be incorporated into t students' grades, he decides to develop a holistic rubric. He identifies the following four attributes on w focus his rubric: estimation, mathematical computation, conclusions, and communication of explanation 2 & 3). Finally, he begins drafting descriptions of the various levels of performance for the observable a (steps 4 & 5). The final rubric for his task appears in Table 4.

Table 4:
Math Performance Task – Scoring Rubric
Data Analysis

Name ____________________________ Date ____________

Score	Description
4	Makes accurate estimations. Uses appropriate mathematical operations with no mistakes. Draws logical conclusions supported by graph. Sound explanations of thinking.
3	Makes good estimations. Uses appropriate mathematical operations with few mistakes. Draws logical conclusions supported by graph. Good explanations of thinking.
2	Attempts estimations, although many inaccurate. Uses inappropriate mathematical operations, but with no mistakes. Draws conclusions not supported by graph. Offers little explanation.
1	Makes inaccurate estimations. Uses inappropriate mathematical operations. Draws no conclusions related to graph. Offers no explanations of thinking.
0	No response/task not attempted.

Example 2:
Subjects - Social Studies; Probability & Statistics
Grade Level(s) - 9 - 12

Mrs. Wolfe is a high school American government teacher. She is beginning a unit on the electoral proc knows from past years that her students sometimes have difficulty with the concepts of sampling and el polling. She decides to give her students a performance assessment so they can demonstrate their levels understanding of these concepts. The main idea that she wants to focus on is that samples (surveys) can accurately predict the viewpoints of an entire population. Specifically, she wants to be able to assess her students on the following instructional objectives:

- Students will collect data using appropriate methods.
- Students will accurately analyze and summarize their data.
- Students will effectively communicate their results. (step 1)

Since the purpose of this performance task is formative in nature, she decides to develop an analytic rub focusing on the following attributes: sampling technique, data collection, statistical analyses, and communication of results (steps 2 & 3). She drafts descriptions of the various levels of performance for observable attributes (steps 4 & 5). The final rubric for this task appears in Table 5.

Table 5:
Performance Task – Scoring Rubric
Population Sampling

Name ____________________ Date ______________

	Beginning 1	Developing 2	Accomplished 3	Exemplary 4	
Sampling Technique	Inappropriate sampling technique used	Appropriate technique used to select sample; major errors in execution	Appropriate technique used to select sample; minor errors in execution	Appropriate technique used to select sample; no errors in procedures	
Survey/ Interview Questions	Inappropriate questions asked to gather needed information	Few pertinent questions asked; data on sample is inadequate	Most pertinent questions asked; data on sample is adequate	All pertinent questions asked; data on sample is complete	
Statistical Analyses	No attempt at summarizing collected data	Attempts analysis of data, but inappropriate procedures	Proper analytical procedures used, but analysis incomplete	All proper analytical procedures used to summarize data	
Communication of Results	Communication of results is incomplete, unorganized, and difficult to follow	Communicates some important information; not organized well enough to support decision	Communicates most of important information; shows support for decision	Communication of results is very thorough; shows insight into how data predicted outcome	
				Total Score	

Resources for Rubrics on the Web

The following is just a partial list of some Web resources for information about and samples of scoring

- "Scoring Rubrics: What, When, & How?" (http://ericae.net/pare/getvn.asp?v=7&n=3). This articl appears in Practical Assessment, Research, & Evaluation and is authored by Barbara M. Moskal. article discusses what rubrics are, and distinguishes between holistic and analytic types. Example additional resources are provided.
- "Performance Assessment-Scoring" (http://www.pgcps.pg.k12.md.us/~elc/scoringtasks.html). Sta Prince George's County (MD) Public Schools have developed a series of pages that provide descr of the steps involved in the design of performance tasks. This particular page provides several rub samples.
- "Rubrics from the Staff Room for Ontario Teachers" (http://www.odyssey.on.ca/~elaine.coxon/rubrics.htm) This site is a collection of literally hundre teacher-developed rubrics for scoring performance tasks. The rubrics are categorized by subject ar type of task. This is a fantastic resource…check it out!
- "Rubistar Rubric Generator" (http://rubistar.4teachers.org/)
- "Teacher Rubric Maker" (http://www.teach-nology.com/web_tools/rubrics/) These two sites hous Web-based rubric generators for teachers. Teachers can customize their own rubrics based on tem on each site. In both cases, rubric templates are organized by subject area and/or type of performa These are wonderful resources for teachers!

References

Airasian, P. W. (2000). *Assessment in the classroom: A concise approach* (2nd ed.). Boston: McGraw-Hill.

BEST COPY AVAILABLE

Airasian, P. W. (2001). *Classroom assessment: Concepts and applications* (4th ed.). Boston: McGraw-Hill.

Chase, C. I. (1999). *Contemporary assessment for educators.* New York: Longman.

Mertler, C. A. (2001). Using performance assessment in your classroom. Unpublished manuscript Bowling Green State University.

Montgomery, K. (2001). *Authentic assessment: A guide for elementary teachers.* New York: Longman.

Moskal, B. M. (2000). Scoring rubrics: what, when, and how?. *Practical Assessment, Research, Evaluation,* 7(3). Available online: http://ericae.net/pare/getvn.asp?v=7&n=3

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merril

Tombari, M. & Borich, G. (1999). *Authentic assessment in the classroom: Applications and practice*. Upper Saddle River, NJ: Merrill.

Trice, A. D. (2000). *A handbook of classroom assessment*. New York: Longman.

Contact information

Craig A. Mertler
Educational Foundations & Inquiry Program
College of Education & Human Development
Bowling Green State University
Bowling Green, OH 43403

mertler@bgnet.bgsu.edu
Phone: 419-372-9357 Fax: 419-372-8265

Descriptors: *Rubrics; Scoring; *Student Evaluation; *Test Construction; *Evaluation Methods; Grades; Grading; *Scoring

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title ▾ [Go]

Copyright 2001, ERIC Clearinghouse on Assessment and Evaluation.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Rudner, Lawrence & Phill Gagne (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, 7(26). Available online: http://ericae.net/pare/getvn.asp?v=7&n=26. This paper has been viewed 913 times since 12/12/01.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs
ERIC/AE Abstracts In Progress

▸ Find articles in ERIC written by
Rudner, Lawrence
Phill Gagne

An Overview of Three Approaches to Scoring Written Essays by Computer

Lawrence Rudner and Phill Gagne
University of Maryland, College Park

It is not surprising that extended-response items, typically short essays, are now an integral part of most large-scale assessments. Extended response items provide an opportunity for students to demonstrate a wide range of skills and knowledge, including higher order thinking skills such as synthesis and analysis. Yet assessing students' writing is one of the most expensive and time-consuming activities for assessment programs. Prompts need to be designed, rubrics created, multiple raters need to be trained, and then the extended responses need to be scored, typically by multiple raters. With different people evaluating different essays, interrater reliability becomes an additional concern in the writing assessment process. Even with rigorous training, differences in the background, training, and experience of the raters can lead to subtle but important differences in grading (Blok & de Glopper, 1992, Rudner, 1992).

Computers and artificial intelligence have been proposed as tools to facilitate the evaluation of student essays. In theory, computer scoring can be faster, reduce costs, increase accuracy and eliminate concerns about rater consistency and fatigue. Further, the computer can quickly rescore materials should the scoring rubric be redefined. This articles describes the three most prominent approaches to essay scoring.

Systems

The most prominent writing assessment programs are

- Project Essay Grade (PEG), introduced by Ellis Page in 1966,
- Intelligent Essay Assessor (IEA), first introduced for essay grading in 1997 by Thomas Landauer and Peter Foltz, and

- E-rater, used by Educational Testing Service (ETS) and developed by Jill Burstein.

Descriptions of these approaches can be found at the web sites listed at the end of this article and in Whittington and Hunt (1999) and Wresch (1993). Other software projects are briefly mentioned in Breland and Lytle (1990), Vetterli and Furedy (1997), and Whissel (1994).

Page uses a regression model with surface features of the text (document length, word length, and punctuation) as the independent variables and the essay score as the dependent variable. Landauer's approach is a factor-analytic model of word co-occurrences which emphasizes essay content. Burstein uses a regression model with content features as the independent variables.

PEG - PEG grades essays predominantly on the basis of writing quality (Page, 1966, 1994). The underlying theory is that there are intrinsic qualities to a person's writing style called trins that need to be measured, analogous to true scores in measurement theory. PEG uses approximations of these variables, called proxes, to measure these underlying traits. Specific attributes of writing style, such as average word length, number of semicolons, and word rarity are examples of proxes that can be measured directly by PEG to generate a grade. For a given sample of essays, human raters grade a large number of essays (100 to 400), and determine values for up to 30 proxes. The grades are then entered as the criterion variable in a regression equation with all of the proxes as predictors, and beta weights are computed for each predictor. For the remaining unscored essays, the values of the proxes are found, and those values are then weighted by the betas from the initial analysis to calculate a score for the essay.

Page has over 30 years of research consistently showing exceptionally high correlations. In one study, Page (1994) analyzed samples of 495 and 599 senior essays from the 1998 and 1990 National Assessment of Educational Progress using responses to a question about a recreation opportunity: whether a city government should spend its recreation money fixing up some abandoned railroad tracks or converting an old warehouse to new uses. With 20 variables, PEG reached multiple Rs as high as .87, close to the apparent reliability of the targeted judge groups.

IEA - First patented in 1989, IEA was designed for indexing documents for information retrieval. The underlying idea is to identify which of several calibration documents are most similar to the new document based on the most specific (i.e., least frequent) index terms. For essays, the average grade on the most similar calibration documents is assigned as the computer-generated score (Landauer, Foltz, Laham, 1998).

With IEA, each calibration document is arranged as a column in a matrix. A list of every relevant content term, defined as a word, sentence, or paragraph, that appears in any of the calibration documents is compiled, and these terms become the matrix rows. The value in a given cell of the matrix is an interaction between the presence of the term in the source and the weight assigned to that term. Terms not present in a source are assigned a cell value of 0 for that column. If a term is present, then the term may be weighted in a variety of ways, including a 1 to indicate that it is present, a tally of the number of times the term appears in the source, or some other weight criterion representative of the importance of the term to the document in which it appears or to the content domain overall.

Each essay to be graded is converted into a column vector, with the essay representing a new source with cell values based on the terms (rows) from the original matrix. A similarity score is then calculated for the essay column vector relative to each column of the rubric matrix. The essay's grade is determined by averaging the similarity scores from a predetermined number of sources with which it is most similar. Their system also provides a great deal of diagnostic and evaluative feedback. As with PEG, Foltz, Kintsch and Landauer (1998) also report high correlations between IEA scores and human scored essays.

E-rater - The Educational Testing Service's Electronic Essay Rater (e-rater) is a sophisticated "Hybrid Feature Technology" that uses syntactic variety, discourse structure (like PEG) and content analysis (like IEA). To measure syntactic variety, e-rater counts the number of complement, subordinate, infinitive, and relative clause and occurrences of modal verbs (would, could) to calculate ratios of these syntactic features per sentence and per essay. For structure analysis, e-rater uses 60 different features, similar to PEG's proxes.

Two indices are created to evaluate the similarity of the target essay's content to the content of calibrated essays. As described by Burstein, et.al (1998), in their *EssayContent* analysis module, the vocabulary of each score category is converted to a single vector whose elements represent the total frequency of each word in the training essays for that holistic score category. The system computes correlations between the vector for a given test essay and the vectors representing the trained categories. The score that is most similar to the test essay is assigned as the evaluation of its content. E-rater's *ArgContent* analysis module is based on the inverse document frequency, like IEA. The word frequency vectors for the score categories are converted to vectors of word weights. Scores on the different components are weighted using regression to predict human grader's scores.

Analysis

Several studies have reported favorably on PEG, IEA, and e-rater. A review of the research on IEA found that its scores typically correlate as well with human raters as the raters do with each other (Chung & O'Neil, 1997). Research on PEG consistently reports relatively high correlations between PEG and human graders relative to correlations between human graders (e.g., Page, Poggio, & Keith, 1997). E-rater was deemed so impressive it is now operational and used to score the General Management Aptitude Test (GMAT). All of the systems return grades that correlate significantly and meaningfully with those of human raters.

Compared to IEA and e-rater, PEG has the advantage of being conceptually simpler and less taxing on computer resources. PEG is also the better choice for evaluating writing style, as IEA returns grades that have literally nothing to do with writing style. IEA and e-rater, however, appear to be the superior choice for grading content, as PEG relies on writing quality to determine grades.

All three of these systems are proprietary and details of the exact process are not generally available. We do not know, for example, what variables are in any model nor their weights. The use of automated essay scoring is also somewhat controversial. A well-written essay about baking a cake could receive a high score if PEG were used to grade essays about causes of the American Civil War. Conceivably, IEA could be tricked into giving a high score to an essay that was a string of relevant words with no sentence structure whatsoever. E-rater appears to overcome some of these criticisms at the expense of being fairly complicated. These criticisms are more problematic for PEG than for IEA and e-rater.

One should not expect perfect accuracy from any automated scoring approaches. The correlation of human ratings on state assessment constructed-response items is typically only .70 - .75. Thus, correlating with human raters as well as human raters correlate with each other is not a very high, nor very meaningful, standard. Because the systems are all based on normative data, the current state of the art does not appear conducive for scoring essays that call for creativity or personal experiences. The greatest chance of success for essay scoring appears to be for long essays that have been calibrated on large numbers of examinees and which have a clear scoring rubric.

Those who are interested in pursuing essay scoring may be interested in the Bayesian Essay Test Scoring sYstem (BETSY), being developed by the author based on the naive Bayes text classification literature (e.g., McCallum and Nigam, 1998). Free software is available for research use.

While recognizing the limitations, perhaps it is time for states and other programs to consider automated scoring services. We don't advocate abolishing human raters. Rather we can envision the use of any of these technologies as a validation tool with each essay scored by one human and by the computer. When the scores differ, the essay would be flagged for a second read. This would be quicker and less expensive than current practice.

We would also like to see retired essay prompts used as instructional tools. The retired essays and grades can be used to calibrate a scoring system. The entire system could then be made available to teachers to help them work with students on writing and high-order skills. The system could also be coupled with a wide range of diagnostic information, such as the information currently available with IEA.

Key web sites

PEG - http://134.68.49.185/pegdemo/ref.asp
IEA - http://www.knowledge-technologies.com/
E-rater - http://www.ets.org/research/erater.html
BETSY - http://ericae.net/betsy/

References and Recommended Reading

Blok, H., & de Glopper, K. (1992). Large-scale writing assessment. In L. Verhoeven (Ed.), J. H. A. L. De Jong (Ed.), *the Construct of Language Proficiency: Applications of Psychological Models to Language Assessment*, pp. 101-111. Amsterdam, Netherlands: John Benjamins Publishing Company.

Breland, H. M., & Lytle, E. G. (1990). Computer-assisted writing skill assessment using WordMAP. ERIC Document Reproduction Service No. ED 317 586.

Burstein, J., K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M.D. Harris (1998). Automated scoring using a hybrid feature identification technique. In the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, August, 1998. Montreal, Canada. Available on-line: http://www.ets.org/reasearch/aclfinal.pdf

Burstein, J. (1999). Quoted in Ott, C. (May 25, 1999). Essay questions. *Salon.* Available online: http://www.salonmag.com/tech/feature/1999/05/25/computer_grading/

Chung, G. K. W. K., & O'Neil, H. F., Jr. (1997). Methodological Approaches to Online Scoring of Essays. ERIC Document Reproduction Service No. ED 418 101.

Fan, D. P., & Shaffer, C. L. (1990). Use of open-ended essays and computer content analysis to survey college students' knowledge of AIDS. *College Health, 38*, 221-229.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, *25*, (2&3), 285-307.

Jones, B. D. (1999). Computer-rated essays in the English composition classroom. *Journal of Educational Computing Research, 20*(2), 169-187.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W, & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

McCallum, A. and K. Nigam (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on "Learning for Text Categorization". Available on-line: http://citeseer.nj.nec.com/mccallum98comparison.html

McCurry, N., & McCurry, A. (1992). Writing assessment for the twenty-first century. *Computer Teacher, 19*, 35-37.

Page, E. B. (1966). Grading essays by computer: Progress report. Notes from the 1966 Invitational Conference on Testing Problems, 87-100.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2), 127-42.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). Computer analysis of student essays: Finding trait differences in the student profile. AERA/NCME Symposium on Grading Essays by Computer.

Rudner, L.M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation*, 3(3). Available online: http://ericae.net/pare/getvn.asp?v=3&n=3.

Vetterli, C. F., & Furedy, J. J. (1997). Correlates of intelligence in computer measured aspects of prose vocabulary: Word length, diversity, and rarity. *Personality and Individual Differences, 22(*6), 933-935.

Whissel, C. (1994). A computer program for the objective analysis of style and emotional connotations of prose: Hemingway, Galsworthy, and Faulkner compared. *Perceptual and Motor Skills, 79*, 815-824.

Whittington, D., & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. *Proceedings of the Third Annual Computer Assisted Assessment Conference,* 207-219. Available online: http://cvu.strath.ac.uk/dave/publications/caa99.html.

Wresch, W. (1993) The Imminence of Grading Essays by Computer - 25 Years Later. *Computers and Composition,* 10(2), 45-58. Available online: http://corax.cwrl.utexas.edu/cac/archiveas/v10/10_2_html/10_2_5_Wresch.html.

Descriptors: Essays; Constructed Response; Scoring; Artifical Intelligence